Automatic Understanding of Unwritten Languages

A thesis presented
by

Oliver Adams

to
School of Computing and Information Systems
in total fulfillment of the requirements
for the degree of
Doctor of Philosophy

The University of Melbourne
Melbourne, Australia
December 2017

Declaration

This is to certify that:

(i) the thesis comprises only my original work towards the PhD except where indicated in the Citations to Previously Published Work;

(ii) due acknowledgement has been made in the text to all other material used;

(iii) the thesis is fewer than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Signed: _____ Date: _____

Thesis advisor(s)
Steven Bird
Trevor Cohn
Graham Neubig

Author
Oliver Adams

Automatic Understanding of Unwritten Languages

Abstract

Many of the world's languages are falling out of use without a written record and minimal linguistic documentation. Language documentation is a slow process and there are an insufficient number of linguists working to ensure the world's languages are documented before they die out. This thesis addresses automatic understanding of unwritten languages in order to perform tasks such as phonemic transcription and bilingual lexicon induction. The automation of such tasks promises to improve the leverage of field linguists and ultimately speed up the language documentation process.

Modelling endangered languages is challenging due to the nature of the available data, which is typically not written text but limited quantities of recorded speech. Manually annotated information in the form of lexicons and grammars is typically also limited. Since the languages are spoken, the most efficient way of sourcing data is to collect speech in the language. Most speakers of endangered languages are bilingual or multilingual, so acquiring spoken translations works to the strength of the speakers. Key approaches described in this thesis make use of bilingual data, in particular translated speech, which consists of segments of endangered language speech paired with translations in a larger language. Such data is important for relating the source language speech with a larger language. Additionally, the application of monolingual phoneme transcription is also explored, since it has direct applicability in more traditional phonemic transcription workflows. The overarching question is this: what can be automatically learnt about the languages with the data we have available, and how can this help automate language documentation?

We first consider translation modelling of accurate phoneme transcriptions. This assumption allows us to investigate the feasibility of phoneme–word translation and the effectiveness of inferring bilingual lexical items from such data in isolation from confounding acoustic factors. A second investigation explores how bilingual lexicons can be used to improve language models, which are crucial components of speech recognition and machine translation systems. In a third set of experiments we remove the assumption of accurate transcriptions and investigate operating in the face of acoustic uncertainty. Experiments in this space demonstrate that translated speech can improve automatic phoneme transcription even without a prior translation model. Finally, we make a step towards further generalisability, exploring acoustic modelling in resource-scarce environments without a lexicon or language model. In particular, we assess the use of automatic phoneme and tone transcription on Yongning Na, a threatened tonal language spoken in south-west China. Beyond quantitative investigation, we report on the use of this method in linguistic documentation of Na. Its effectiveness has led to its incorporation into the language documentation workflow for Na.

Citations to Previously Published Work

The work in this thesis is entirely original except where explicitly stated otherwise.

Large portions of Chapter 3 have appeared in the following paper:

> Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird (2015) Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions, in *Proceedings of the 12th International Workshop on Spoken Language Technologies (IWSLT)*, Da Nang, Vietnam. pp. 248–255.

Large portions of Chapter 4 have appeared in the following paper:

> Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, Trevor Cohn (2017) Cross-lingual word embeddings for low-resource language modeling, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

Large portions of Chapter 5 have appeared in the following papers:

> Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird (2016) Learning a translation model from word lattices, in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, California, USA.

> Oliver Adams, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Do Truong, Satoshi Nakamura (2016) Learning a lexicon and translation model from phoneme lattices, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA. pp. 2377–2382.

Large portions of Chapter 6 have appeared in the following papers:

Oliver Adams, Trevor Cohn, Graham Neubig, Alexis Michaud (2017) Phonemic transcription of low-resource tonal languages, in *Proceedings of the Australasian Language Technology Association Workshop 2017 (ALTA)*, Brisbane, Australia. pp. 53–60.

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, Alexis Michaud (2018) Evaluating phonemic transcription of low-resource tonal languages for language documentation in *Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan. (To appear).

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost I thank my parents David and Joan, who have given me a very fortunate life. You have always believed and trusted in me, and given me so much support beyond what can be expected on the basis solely of kinship. To that I add my sister Sophie, who has always been a great source of intellectual stimulation in the form of deep and meaningful conversations.

I thank my supervisors: Steven Bird for giving me the opportunity to do the PhD and for introducing me to this weird and wonderful world of research; Trevor Cohn for subsequently coming along, dumping a whole lot of ideas in my head and sharing a few laughs with me while at it; and Graham Neubig for being so willing to spend time communicating with and helping me via email for years despite no obligation to, and for giving me the chance to visit NAIST for a few months, which was the most productive period of my life thus far.

Alexis Michaud has been a fantastic collaborator, so helpful and enthusiastic in our endeavours. It is his contagious enthusiasm that has been a source of much of my own drive in the past year. May our long-winded email exchanges continue into the future!

I've been lucky enough to have interned with people who I really clicked with. Shourya Roy, my mentor at Xerox Research Center India, was very supportive while at the same time giving me much freedom in the research project. The other interns at XRCI were great company and were so willing to help me in times of need: Archana Sahu, Dhawal Joharapurkar, Raksha Sharma, Nirmal Jith, and Prajna Upadhyay to name a few.

At IBM Research Australia I had the great fortune of being mentored by Adam Makarucha, who is an all-round good bloke. Aside from making me feel so welcome and being so supportive, our interesting and wide-ranging lunchtime chats with another great guy, George Yiapanis, were the perfect way to break up the many productive days I had during the internship.

At NAIST, under Neubig-sensei's wing, I met many good people and I reflect on those days fondly. While my stay was largely an exercise in research monasticism, interactions with a number of key players made it all the better. The senior figures, Nakamura-sensei, Sakti-sensei and Matsuda-san, made me feel so welcome, while the

student crew were great company: Michael Heck, Nunu Lubis, Philip Arthur, Do Truong, Patrick Lumbantobing, Masahiro Mizukami, Makoto Morishita, Matthias Sperber (briefly!) and many more.

During my time at Melbourne Uni I've been fortunate to meet and get to know so many great people: Florian Hanke, who shared a cubicle with me in the first couple years and made for a solid coffee buddy. Florian taught me much and helped me substantially in the Aikuma programming endeavour while also facilitating such adventures as: the failed summitting of Mt. Kosciuszko and a weekend in Lisbon; Long Duong who, aside from being an inspiration in the academic arena, made for enjoyable tennis sessions, impressing me with rapidly gained proficiency in the game; Ned Letcher for all the common interests and good conversation; Nitika Marthur for probing, thoughtful conversation, and so many more who shaped me and my experience in 8.19: Meladel Mistica, Marco Lui, Richard Fothergill, Bahar Salehi, Jey Han Lau (also for tennis and valuable ideas at IBM), Li Wang, Afshin Rahimi, Doris Hoogeveen, Miji Choi, Katya Vylomova, Fei Liu, Yitong Li, Aili Shen, Mohammad Oloomi and so many more who I could say good things about but I'll save the trees. More senior people include Yvette Graham, Paul Cook, Joel Nothman, Julian Brooke, Meng Fang, and now Bahar again! Tim Baldwin is the leader of the NLP group and a role model I hold in very high esteem. Among other things I must thank him for, I realise as I write this that he has had an instrumental role in gathering together the above people that have made for such a hospitable research group.

There is a seemlingly endless list of people from elsewhere who have helped me along the way. These include Marco Vetter, the annotator of lexical entries in Chapter 3; Antonios Anastasopoulos, who drove down from Indiana to Pittsburgh for an impromptu collaboration; people from CMU such as Florian Metze and Alan Black for imparting their wisdom during my short stay there; and Felix Stahlberg, who planted the idea of phoneme recognition of Yongning Na in my head and pointed me in the direction of Alexis Michaud. Thank you everybody!

Finally, I thank the thesis examiners who took the time to thoroughly read this thesis. Their thoughtful and clear advice has definitely helped to improve the document.

Chapter 1

Introduction

Rapid extinction of the world's underdocumented, unwritten languages motivates the incorporation of automation into language documentation work. Two central procedures in the language documentation process are phonemic transcription and the construction of bilingual lexicons. This thesis addresses *automatic understanding* of endangered language speech, which has the potential to provide linguists with greater leverage in performing these tasks. Beyond its promise in helping to understand and preserve the world's linguistic heritage, the advancement of computational methods in this space has the potential to improve language technology for low-resource languages that are not threatened or endangered, but instead have many speakers who would stand to benefit from such technology.

This chapter discusses the phenomenon of language extinction and language shift, what people are doing about it, and how technology is changing the documentation process and the nature of the data being collected. The aim and scope of the thesis is then presented, along with a summary of contributions.

1.1 Motivation

1.1.1 Language Extinction and Documentation

The majority of the world's languages are losing speakers and it is predicted that between 50 and 90 per cent of these languages will become extinct in the next 100 years (Krauss 1992; Miyaoka *et al.* 2007; Austin and Sallabank 2011). About half of the world's approximately 7,000 languages have no orthography and thus no written record (Lewis *et al.* 2015). Unless these languages are documented before they fall out of use, much linguistic, cultural and anthropological information will be lost forever.

Language extinction motivates field linguists to engage in the documentation of languages so that a record of the language may be kept for posterity. Crowley (2007) argues that since the field of linguistics seeks to answer questions about the nature of human languages in order to establish what phenomena can and cannot occur, documentation of these languages is very important. This cannot be done by investigating only large languages, as that would be neglecting consideration of the rich linguistic diversity of human language. However, with such limited documentation of most of the world's smaller languages, it is often impossible to conclusively answer important questions about how languages work in general.

Moving beyond the study of linguistics, language mediates knowledge about groups of humans, their history and the importance of certain concepts in different cultures. Language is the means with which almost all knowledge is captured and shared. Therefore, without documentation, much knowledge held by the speakers of an unwritten language dies when they do. In many cases however, languages do not *die* but undergo a process of *language shift*. In language shift, intergenerational transfer of the language does occur, but information about entire genres (such as traditional rituals or knowledge of the land) is not passed on to younger speakers. In such cases, much knowledge dies even when the language continues to be spoken.

Language extinction is not just something that may occur, it is happening right now. Every day, speakers of undocumented languages are being lost. In the 20th century many languages died out. Ethnologue reports 367 languages that have be-

come extinct since 1950 at an average rate of 6 languages per year (Lewis *et al.* 2015). Because of the motivation for people—particularly the young—to shift to larger languages, it is reasonable to expect that this rate of extinction will increase.

There may be little that can be done to prevent many languages from falling out of use. People switch to larger languages that are more advantageous for them for employment and some speakers consider it progress to do so (Harrison 2008). In some places such as in Tanzania, it is argued that a shift to a unified language is a step away from tribalism which is seen as a threat to development (Ladefoged 1992). While the value of keeping a language alive is not universally accepted, this is not the main goal of language documentation. With adequate linguistic documentation there is the potential to save much of this valuable information, even when the language is no longer spoken. Incorporating automation into the documentation process has the potential to reduce the time it takes to achieve such documentation of a language.

1.1.2 The Changing Nature of Language Documentation

Traditionally, language documentation involves the linguist travelling to remote communities for one-on-one elicitation of speech data from speakers of threatened languages, followed by manual analysis to produce text collections, lexicons and grammars of their languages. However, this process is slow and there are a limited number of linguists engaged in such work. Given the estimated rate of language extinction, it's clear that the current rate of collection is insufficient and that a radical speedup is needed in order to adequately document most of the world's languages before they die out.

An important part of language documentation is the acquisition of bilingual data relating the threatened or endangered language to a larger regional language or international language. Such data is rich in information, even in small quantities. (Note that throughout this thesis we use the terms "low-resource," "threatened," "endangered," "larger," and "high-resource" languages. How best to define these terms is open to debate, but for the scenario relevant to this thesis we generally use "low-resource" to refer to those languages categorized as between 6b and 10 on the EGIDS

scale (Lewis and Simons 2010), while languages between 0 and 6a are considered "larger." Category 0 languages are "high" resource.)

The proliferation of cheap mobile phones is creating new opportunities for documenting languages in a manner more efficient than traditional approaches (Hughes *et al.* 2010; Reiman 2010; De Vries *et al.* 2014; Bird *et al.* 2014b; Bettinson and Bird 2016). *Aikuma*[1] and an extended version, *Lig-Aikuma*[2] (Blachon *et al.* 2016) used in the Breaking the Unwritten Language Barrier (BULB) project (Adda *et al.* 2016) (discussed in Chapter 2), are two such apps that aim to provide field linguists with greater leverage in eliciting speech data through the use of a crowdsourcing model whereby smartphones can be distributed amongst speakers of a language for elicitation of source speech with minimal supervision. Since most speakers of endangered languages are bilingual, Aikuma aims to elicit bilingual speech aligned in segments between the endangered language and the larger language, the latter of the two being a language that can be more reliably transcribed. *Zahwa*[3] is a similar application that supports spoken translation, originally intended for describing food preparation with accompanying pictures, but appropriate for documenting and translating any procedural knowledge (Bettinson and Bird 2016).

Though we will discuss the current breadth of such apps in the next chapter, with more and more people using electronics as a part of their daily routine, it's likely that further methods of performing language documentation will appear. There are three aspects of these methods that will likely continue to be popularised: (a) less involvement from linguists per recording, (b) less manual linguistic annotation per recording, and (c) increased leveraging of bilingual speech. (a) and (b) are a natural result of the attempt to collect more data from more speakers without a proportional increase in the numbers of trained linguists. Less involvement from the linguist also has the advantage that it makes it easier to collect language as it is spoken in natural, often colloquial, contexts. (c) capitalizes on the very cause of language extinction: the movement of people away from their native language towards larger languages,

[1] `aikuma.org`
[2] `lig-aikuma.imag.fr`
[3] `zahwa.aikuma.org`

which means a lot of speakers are bilingual. Working to the strength of bilingual speakers is important as it is easier to elicit spoken translations than transcriptions when a language has no standard orthography. It is also important to relate the source speech to a larger language through some form of bilingual mapping. However, in the context of language shift, younger bilinguals may not always be able to easily translate content from older monolinguals due to a significant loss of vocabulary in the process of intergenerational transfer. This is an additional challenge and adds to the pressure for translations to be made now while the original speakers are available for consultation.

But can't we just wait for the rapid adoption of mobile phones and computers to perform the language documentation for us? Since the movement towards such technology usually occurs with a switch to communication in a larger language, it is likely that little information from the endangered language will ultimately be captured unless there is an explicit push to do so.

1.1.3 A Different Kind of Data

The data resulting from this changing face of language documentation has specific features that distinguish it from the data used in most natural language processing (NLP) research. The data primarily addressed in this thesis is small amounts of speech data in an endangered language with translations into a larger language. Effective processing of this data to model the language and speed up the work of the linguist is a challenge. We now describe the key distinguishing properties of such data.

Limited Quantity Though smartphone-based rapid language documentation offers promise that such data can be acquired more quickly than it has been previously, the amount of such data available will always be much less than what is available for natural language processing of larger languages. Expectations of the amount of data available should thus be kept conservative for the purposes of our investigation so that any conclusions drawn are generalizable to as many low-resource language documentation contexts as possible.

For example:

- In Tembé, Pará State, Brazil, Bird *et al.* (2014a) collect 2 hours of source audio with 35 minutes of this audio orally translated over a 4 day visit.
- Nhengatu, Amazonas State, Brazil, the same authors collect 2.5 hours of source audio and orally translate 1 hour over a 3 day visit.
- In Brazzaville, Congo, Blachon *et al.* (2016) collected 48 hours of Mboshi speech using Lig-Aikuma in two 1-month field trips. Of this, 5.5 hours were spoken translations elicited from French text (Godard *et al.* 2017).

If we assume the rule of thumb that 1 hour of audio equals around 1,000 sentences (Cieri and Liberman 2006; Bird and Chiang 2012; Bird *et al.* 2014b), then in these specific cases at most a few thousand sentences of bilingual audio are collected.

Bilingual It is typical in traditional language documentation to create glosses and bilingual lexicons. Since speakers of endangered languages are frequently bilingual in order to communicate with a wider regional community, the data collected harnesses this extra source of information. The investigations of this thesis concern data that was originally produced in the endangered source language, with some subset translated into a larger regional language, such as Brazilian Portuguese in Brazil.

Audio Most endangered languages do not have a standardized orthography, making written documentation in the language difficult and encouraging the emphasis of documentation to be on speech. Regardless, collection of spoken recordings has three main advantages:

1. It is faster than collecting text. It is also a precondition for transcription. Phonemic transcription is a slow process, taking a trained linguist roughly 1 hour to transcribe 1 minute of speech (Do *et al.* 2014a), or 50 to 100 hours per recorded hour (Cavar *et al.* 2016). Focusing only on capturing audio and its subsequent processing avoids the slow process of manual transcription.

2. It captures significant amounts of information about intonation, articulation and accents, that can be lost in transcription.

3. It captures language as it naturally occurs: in fluent conversation, which deviates from written speech in multiple dimensions (Redeker 1984). It is therefore data that is more representative of the language.

However, the use of spoken recordings means the data is subject to the challenges of automatic speech recognition. The fluent and conversational nature of speech makes it difficult to transcribe on account of its variable features such as prosody, coarticulation and disfluencies. Moreover, when automatically transcribing phonemes without the aid of a language model, phoneme error rates are high.

Though the data consists of spoken recordings, we assume in much of the thesis that the target side is a larger language that can be efficiently transcribed. While the issue of transcription still remains on the target side, and may be complicated by the bilingual speakers speaking a non-standard or accented form of the target language (which may require an additional *respeaking* step), the prospects of efficient, accurate and scalable manual or automatic transcription with a standard orthography in this target language are far greater than in the source language. Such transcription may be automatic, since breakthroughs in automatic speech recognition have made state-of-the-art systems effective for large languages. In addition, manual transcription (perhaps via crowdsourcing) is also more feasible for large languages, due to the large number of speakers.

Limited prior linguistic information The extensive body of linguistic research on English can inform the approaches used in natural language processing systems for English. Methods for automation in endangered language documentation has much less information to draw from. That said, we can assume some linguistic information. It is reasonable to expect a linguist can determine the phoneme inventory of a language in a comparatively short amount of time. It might also be reasonable to assume elicited spoken recordings of common words can be acquired. The small amount of prior information means that it must be used as effectively as possible.

1.1.4 Why Pursue Automatic Understanding of Unwritten Languages?

There are a number of reasons motivating development of speech recognition and translation tools for endangered languages. Firstly, the availability of such technology may help to speed up language documentation by providing linguists with greater leverage in data collection. If linguists are bogged down less in the time-consuming task of phonemic transcription, that effort can be spent on expanding the scope of their research and their reach to other languages. Additionally, the linguist may be able to spend more time on linguistically meaningful dialogue with the consultants when the burden of manual data entry is reduced (as we see in §6.5). Although collection of primary speech data is a key issue in language documentation, this secondary processing is a bottleneck in the standard documentary linguistics workflow. Linguists accumulate considerable amounts of speech, but do not transcribe and translate it all. There is a risk that untranscribed recordings could end up as "data graveyards" (Himmelmann 2006:4,12-13), and as a result, there is a need for "devising better ways for linguists to do their work" (Thieberger 2016:92). "For example, out of the 137 unrestricted collections in the Archive of the Indigenous Languages of Latin America, about half (49%) contain no transcriptions at all, and only 7% are fully transcribed" (Anastasopoulos *et al.* 2017). In addition to speeding up this process, the quality and experience of the linguist's work may improve since the automatic methods may highlight phenomena the linguist might otherwise overlook, such as interesting phonetic and phonemic facts (see Chapter 6) and possible bilingual lexical entries.

A second motivation is that making language technology available for threatened languages may help in revitalization of these languages. In cases where the community is interested in revitalization, the existence of such language technology may make the languages be perceived as more relevant by the children of native speakers, who might otherwise not carry on the tradition of the language. The benefits of this are not purely sentimental: slowing the rate of language extinction and possibly even revitalizing languages affords communities more time for the languages to be documented, meaning fewer languages will fall out of use without being documented. In

this sense, documentation and revitalization are symbiotic. Both the data collected and the the technology developed for language documentation may facilitate revitalization, furthering the coverage of the documentation. The availability of machine translation technology for low-resource languages also has the potential to make them more relevant by translating educational resources available only in larger languages into the smaller languages. However this objective raises its own problems for NLP. Firstly, there is a domain mismatch between translated heritage materials from the source language used to train the models and the genre of what is to be translated from the larger language, which may be information such as educational material or Wikipedia content. Secondly, the lack of a large amount of source language text means language model quality in the source language is likely to be poor, limiting machine translation (MT) performance.

A third motivation is that of incident-response, the focus of the DARPA LORELEI project.[4] There is value in having language technology available "in the context of a rapidly emerging and quickly evolving situation like a natural disaster or disease outbreak" (Strassel and Tracey 2016). Humanitarian assistance and disaster relief can benefit from technology such as translation tools to help communication during a time of crisis, for example in the 2010 Haiti earthquake, which highlighted this need (Besacier *et al.* 2014).

Finally, *constraints breed creativity.* Limits on the available data force innovation in modelling approaches, and such approaches are useful beyond the arena of endangered unwritten languages. There exist many low-resource languages that are not endangered. These languages are characterized by having many speakers (perhaps millions), robust intergenerational transfer, an established main dialect, and some web presence, but still lacking the data available for state-of-the-art language technology. Research on language technology for endangered languages can inform approaches for such low-resource languages, which has the potential to positively impact millions of speakers. Moreover, there are data sparsity issues even for large languages when we consider issues arising from morphology and constraints on the

[4] `www.darpa.mil/program/low-resource-languages-for-emergent-incidents`

domain. As a result, advances in this space may find applicability elsewhere in NLP of high-resource languages.

1.2 Aim and Scope

The aim of this thesis is to progress towards effective semi-automated language documentation by modelling and making best use of available data, such as monolingual and bilingual speech. Automatic understanding of languages through transcription and translation modelling may efficiently guide future documentation of those languages and engage more linguists by helping to speed up very slow work such as manual phonemic transcription. Furthermore, investigating how to deal with very low-resource languages will help when addressing languages for which there are many speakers, but nevertheless have limited resources, by prompting focus on improved modelling of language.

The scope of this PhD is limited to the tasks of bilingual lexicon induction, language modelling, and phonemic transcription. In Chapters 3 and 5, the input data is speech or a phonemic representation, along with an orthographic translation in another language. Orthographic translations are assumed since the target language is typically a large language for which manual or automatic transcription is efficient. (It is often the case that the variety of the larger language spoken by the bilingual speaker is far from standard, which may motivate a further *respeaking* of the translation as supported by Lig-Aikuma (Blachon *et al.* 2016)). In Chapter 4 we consider the use of bilingual lexicons and large amounts of text data in another language in order to enable transfer learning when we have limited transcribed data with which to train language models. Finally, in Chapter 6 limited monolingual source speech and phonemic transcriptions are used to train acoustic models for phoneme and tone prediction.

One consistent assumption about available data that holds across all the research described in this thesis is the availability of a phonemic inventory (and where relevant to the language, a tone inventory) identified by a linguist. This is a reasonable requirement, as eliciting a phonemic inventory scales well and is 'constant' with respect

to the amount of speech gathered, unlike transcription.

1.2.1 Research Questions

The key aim is to determine what can be automatically learnt about the languages with the data we have available, and how this can help automate language documentation. This goal is decomposed into a collection of research questions:

A. Phoneme translation modelling 1. How accurate is translation modelling of unsegmented phonemic transcriptions? 2. How do different translation models for this task compare in bilingual lexicon induction?

B. Speech translation modelling 1. How can translation models be learnt from speech, and 2. can these be used to improve speech recognition and automatic phonemic transcription?

C. Cross-lingual language modelling 1. How can other bilingual resources, such as lexicons, be used to transfer information from a high-resource language to a low-resource language? 2. Can such approaches be used to improve language modelling, which is useful to speech recognition and machine translation?

D. Usefulness for the linguist and tonal languages Many languages we would like to automatically transcribe are tonal, and tonal transcription is an important process in the linguist's workflow. 1. How well can we predict tones for richly tonal languages? 2. How does phoneme and tonal prediction fit into a linguist's actual linguistic workflow?

E. Scaling For the methods used in answering the above questions, we also want to answer the question of how well performance scales with training data size. We're particularly interested in whether such approaches can be usefully applied to small amounts of data to give linguists and computer scientists a sense of how much data is required for methods to be effective.

1.3 Overview of the Contributions

The contributions of this thesis take a step towards answering the above questions. The first contribution of this thesis is an investigation into translation modelling of unsegmented sequences of phonemes for the purpose of bilingual lexicon induction (Chapter 3), addressing questions A1 and A2. This includes end-to-end machine translation experiments as well as a bilingual lexicon induction experiment focusing on the quality of the inferred lexemes. Traditionally, translation modelling has been trained on ample data segmented at the word-level. This contribution involves an evaluation of various methods on small amounts of unsegmented phonemic data, with results indicating that translation modelling can be effective even when training data amounts to as little as 1,000 bilingual sentences of unsegmented phonemic text: an amount feasible for accurate manual transcription. Work here highlights that bilingual lexicon induction, measured by assessing which entries are corroborated by established lexicons, is not a good intrinsic measure of the method's performance. Moreover, linguists tend to create these in the process of phonemic transcription, limiting the applicability of such bilingual lexicon induction techniques.

In light of the availability of lexicons for many languages for which speech recognition and machine translation systems do not exist, the next contribution explores the use of such lexicons for the purposes of transfer learning (Chapter 4), addressing questions C1 and C2. This work demonstrates that the use of bilingual lexicons can improve language models, a key component in speech recognition and machine translation systems, by using them to tap into cross-lingual distributional information. This is demonstrated by initializing neural network language models with cross-lingual word embeddings.

The first two contributions explore scenarios where transcribed speech in a source language is available. The third contribution (Chapter 5) instead explores how translated speech, which is easier to collect than phonemic transcriptions, can be used to improve automatic phoneme transcription. This addresses questions B1 and B2, while giving deeper insight into A1. First, an approach for resolving acoustic modelling errors is explored that relies on the concept of phoneme equivalence classes. In

light of this model's flaws, we proceed to investigate a model that jointly segments phoneme lattices and aligns them with an orthographic translation while learning a lexicon and translation model. Results demonstrate that translations of speech can help resolve errors in phoneme recognition even when there is no prior translation model or bilingual lexicon relating the languages at all. A framework is laid out which can be extended on for further modular improvement.

The contributions described thus far assume either accurate phonemic transcriptions, or the availability of an acoustic model. The final contribution (Chapter 6) addresses the requirement of an acoustic model for automatic phoneme transcription of Yongning Na, a minority language of Yunnan, China, for which there is immediate practical benefit in language documentation work. Addressing questions D1 and D2, we conduct an investigation into acoustic modelling as a way to make automatic phoneme transcription feasible for languages with extremely low amounts of training data. Importantly, we address phonemic transcription of tones, since many languages are tonal and transcription of tones is important in the language documentation. A key takehome is that useful phoneme recognition accuracies can be obtained when trained on small quantities of transcribed data and that such automated transcripts can serve as a useful "canvas," partially automating the work of the linguist. This is encouraging, as we can tell linguists that as little as 30 minutes of transcription work may be adequate to build a training set which can then be used to aid in subsequent transcription. Such automated transcription has now been incorporated into the linguist's workflow for documentation of Na.

The progression of the chapters can be viewed as progressively removing simplifying assumptions about the data while addressing pieces in an automatic phonemic transcription and lexicon induction pipeline: language models, translation models, and acoustic models. In Chapter 3, we assume large quantities of data and error-free phonemic transcriptions in order to assess translation modelling. We then remove the assumption of large quantities of data in order to assess bilingual lexicon induction from limited phonemic transcriptions without word segmentation. Chapter 4 also assumes the availability of correct transcriptions, and bilingual lexicons are used for the purposes of training language models. In Chapter 5 we remove the assumption

of error-free transcriptions. We first do this by simulating errors using two different methods, before using the output of actual acoustic models. In finally addressing acoustic modelling in Chapter 6, we complete exploration of the final piece in the experimental pipeline, dropping the assumption of an acoustic model trained with generous amounts of data. In Chapter 7, we conclude with a discussion of the main findings, limitations of the work and promising avenues for future work.

Chapter 2

Background

The promise of being able to efficiently collect bilingual speech is a key motivation for the contributions of this thesis, and so we begin with a brief survey of work on the acquisition of such input data. Modelling this data requires us to draw on knowledge in both the fields of automatic speech recognition (ASR) and machine translation (MT). This background chapter will provide a lay of the land, giving an overview of relevant work and ideas in MT and ASR before considering the intersection of these two areas of research as they relate to and motivate the subsequent work in the thesis.

This chapter assumes some familiarity with:

- General machine learning concepts (such as the bias/variance tradeoff and the curse of dimensionality) and architectures (particularly neural network models such as feed-forward, recurrent and convolutional architectures).
- Machine translation: word and phrase level alignment and translation, evaluation with BLEU scores. See Koehn (2009) for an accessible reference.
- Speech recognition: traditional approaches based on hidden Markov models. See Jurafsky and Martin (2009) for an accessible reference.
- Linguistic concepts such as parts of speech, derivational and inflectional morphology.

2.1 Data Acquisition

Methods central to this thesis are premised on the acquisition of source language speech paired with translations (spoken or written) in a larger language. Such data is collected in traditional language documentation workflows (Hanke 2017), but these workflows typically entail a slow process of one-on-one work between a linguist and native speaker. This involves creating a phonemic transcription of the primary speech data and relating it to a larger language through the process of glossing and construction of a lexicon, perhaps using a tool such as FLEx.[1]

There have been recent efforts to speed up the language documentation process by harnessing cheap mobile devices to collect speech and spoken translations, as well as fostering collaborative involvement of speakers. These approaches are often based on the idea of greater native speaker involvement in the language documentation process, thereby empowering them and making them custodians of their language rather than the linguists. The involvement of more people in the documentation process may help to speed up progress in documenting the world's languages. At the same time, less direct curation of materials by linguists raises the question of how best to process such data and conduct quality control.

Reiman (2010) proposed the method of basic oral language documentation (BOLD) in order to sidestep the issue of written documentation. This approach involves oral annotation including 'oral transcription:' careful respeaking of a source recording, and 'phrasal translation.' Motivated by the BOLD approach, one collection of work in this space are approaches based on the Aikuma tool (Bird *et al.* 2014b), which is an android app designed to facilitate the collection of recordings of endangered language speech and, importantly, spoken translations in a larger language. Since speakers of endangered languages are frequently bilingual or multilingual, this works to the strength of the speakers. The Aikuma app has had preliminary deployments in Papua New Guinea, Nepal and Brazil (Hanke and Bird 2013; Bird *et al.* 2014a). The Aikuma app was primarily designed for collection of spontaneous speech, as opposed to linguist-driven elicitation of words, phonological and

[1] `software.sil.org/fieldworks/`

grammatical constructs. This encourages the collection of material that is a more faithful representation of the language and what the speakers care about, while at the same time requiring less of the linguist. A positive side effect of this is that the observer effect is reduced: materials are more naturalistic when a foreign linguist is not there holding a microphone.

An extended version of the application, Lig-Aikuma (Blachon *et al.* 2016) has since superseded Aikuma. This application adds several extensions including an elicitation mode. Lig-Aikuma has been used by the *Breaking the Unwritten Language Barrier* (BULB) project (Adda *et al.* 2016) to collect Mboshi speech in Congo and Fongbe speech in Benin through speech elicitation from translated reference sentences (Adda *et al.* 2016; Laleye *et al.* 2016; Godard *et al.* 2017).

Beyond the Aikuma apps, there is other work on smartphone apps to collect speech data. Examples include the app of Hughes *et al.* (2010) for elicitation of speech. Another elicitation-based application is Woefzela (De Vries *et al.* 2011; De Vries *et al.* 2014), the focus of which is to collect speech data from prompts for acoustic model training. However, both of these applications target collection of speech in the same language as the prompt, requiring some textual data beforehand.

The development of applications for language documentation collection is a growing area of work. For further reading, Bettinson and Bird (2016) and Bird (2017) discuss considerations in the development of language documentation applications more generally.

2.2 Modelling Translation

In this section we discuss literature relating to modelling the relationship between words in two or more different languages. The vast majority of work in this space has been focused on bilingual text to address machine translation, where input text in a source language is converted to text in a target language.

Machine translation is a challenging task even for humans, who are the champions of language. There are a number of reasons for this. The relationship between words in languages is not one-to-one: some words may not have a equivalent word

in another language, encouraging either an inaccurate translation or a more elaborate paraphrasing. When there are single-word translations, they often have subtly different meanings in certain contexts. Languages typically have many word tokens which are polysemous, relying on contextual information for translation (sometimes rich background context that a computer cannot yet incorporate into modelling). Furthermore, languages vary greatly in syntax. These issues, along with the effectively unbounded inventory result in problems such as indirect associations (Melamed 1996), where source-language words are given substantial translation probabilities for target-language words that frequently co-occur but are not translations. Such issues make machine learning of effective translation models given parallel data a difficult challenge in artificial intelligence.

This section reviews traditional models of word alignment, recent advances in MT using neural architectures, and approaches addressing data sparsity in MT that are relevant to low-resource translation modelling. Beyond the task of MT, this section also discusses work in related tasks including bilingual lexicon induction and the learning and use of cross-lingual word embeddings.

2.2.1 Traditional Word Alignment

Traditional MT systems from the 1960s until the 1990s relied heavily on linguistically motivated rules (Koehn 2009). Increased computational power and access to corpora catalysed a shift towards statistical methods for machine translation, starting in the late 1980s and gaining momentum through the 1990s. These models are built on large quantities of data and, rather than being governed by handcrafted rules, learn patterns of translation from the data.

In a pivotal paper, Brown *et al.* (1993) introduced the so-called "IBM models" for word-based MT. Given a corpus of text aligned only at the sentence level, the IBM models describe lexical translation probabilities of foreign words given an English word, and alignments of words within these training sentences. In the most basic model, IBM model 1, all possible word orderings have an equal probability and the only parameters the model stores are the lexical translation probabilities. The

subsequent IBM models in the paper build upon this first one to include information such as parameters describing likely word re-orderings and how many English words correspond to a given foreign word. These models play an important role as baselines for the translation modelling in Chapters 3 and 5. As well as the IBM models, other approaches to statistical word alignment arose and gained a foothold, among which is the notable use of hidden Markov models for capturing locality of alignments and allowing the use of well-established statistical methods for training and decoding (Vogel *et al.* 1996; Deng and Byrne 2008).

The parameters of the IBM models 1 and 2 are typically learnt using the *expectation maximization* algorithm. This algorithm uses an iterative two step process whereby the lexical translation parameters of the model are used to predict word alignments (the expectation step) before these word alignments are used to re-estimate the parameters (the maximization step). By iteratively finding the best word alignments and re-estimating the translation model parameters, the model parameters converge to a local optimum that describes the corpora with the highest probability. IBM models 3 and onwards are intractable for expectation maximization and so a hill climbing algorithm is used instead. More recent work in parameter estimation for the IBM models include *Bayesian* methods whereby a distribution over model parameters is instead expressed (DeNero *et al.* 2008; Mermer and Saraçlar 2011; Mermer *et al.* 2013; Li *et al.* 2013). Such approaches can help avoid degenerate solutions and the overfitting that is associated with point estimates of the model parameters found in maximum-likelihood estimation approaches.

2.2.2 What is the Right Granularity for Translation Units?

The translation parameters of the IBM models are at the word level. This is limiting, since the relationships between words and translations may in practice be one-to-many or many-to-many. On the other hand, varying inflection and agglomerative morphology suggest that sub-word translation units may sometimes be more appropriate.

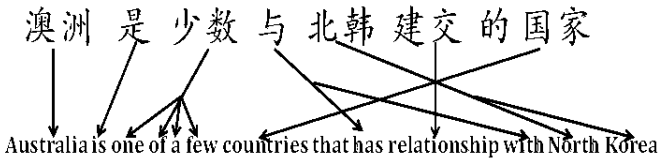


Figure 2.1: Word alignments that motivate multi-word translation units. (Illustration created by Qin Gao, 2008 and released into the Public Domain; example sentence taken from Chiang (2005)).

Phrase Alignment

In order to more accurately model translation of larger collections of words, approaches such as n-gram machine translation (Mariño *et al.* 2006) were introduced. In particular, *phrase-based machine translation* (Koehn *et al.* 2003) became popular for its ability to handle groups of words and local reorderings within them. In this approach, heuristic methods are used to extract phrase alignments from word alignments in order to assign scores to translations of groups of words. This helps resolve alignment ambiguities within the phrases, facilitating a one-to-one mapping between the source and target sentences at the phrase level. Consider how the Chinese–English parallel sentence in Figure 2.1 motivates the use of translation of multi-word units.

A related issue with word-based MT is that multi-word sequences are often not decomposable into translations of the constituent words. Idioms are a clear example of this. Consider the translation of a German idiom (describing a person perceived to be crazy) into an English equivalent:

German:	er	hat	nicht	alle	Tassen	im	Schrank
Gloss:	he	has	not	all	cups	in the	cupboard
English:				‘he’s a fruitcake’			

In the translation between these idioms only the word ‘he’ can be literally translated. This issue is inadequately addressed by heuristic phrase extraction methods since they are based on word-level alignments which ineffectively model such phenomena unless there are enough training examples. Thus while phrase-based MT

was the most popular framework in machine translation until the arrival of neural MT, other machine translation paradigms existed to address this problem of non-decompositionality, among other problems such as limited reordering ability. These include trees (Yamada and Knight 2001; Neubig 2013) and grammars (Chiang 2007). Wu (1997) introduce *inversion transduction grammars* (ITGs), a grammar formalism that permits the vast majority of sentence-level reorderings to be captured in a simple binary synchronous grammar, which have been used as the basis for a variety of phrase alignment models, both with maximum-likelihood and Bayesian inference (Cherry and Lin 2007; Zhang *et al.* 2008; Blunsom *et al.* 2009; Cohn and Haffari 2013). Neubig *et al.* (2011b) contribute one such method, with an important distinguishing feature that the generative story of each parallel sentence involves the model attempting to generate a phrase pair at each branch in the tree, backing off to using a non-terminal distribution. This allows for phrase translations of varying granularities to be modelled without decomposition into tokens. This contrasts with previous ITG approaches that require heuristics to construct larger phrases, as phrase-based MT approaches do. We explore the use of heuristic phrase extraction versus a Bayesian ITG approach for translation modelling and lexicon induction in Chapter 3.

Character-Based Alignment and Sub-Word Neural Machine Translation

While the shift of phrase-based statistical MT was largely based on the idea that word alignments are often too fine-grained to capture larger units appropriate translation, it is also the case that they are limited in that they often are not fine-grained enough. Consider how decomposition of a word into sub-word units such as morphemes reveals information that could be useful in translation. *Character-based MT* addresses this idea in the other extreme, using characters as the unit of translation. Earlier approaches to character-based MT focused on similar languages such as Castilian and Catalan (Vilar *et al.* 2007), Norwegian and Swedish (Tiedemann 2009), and Macedonian and Bulgarian (Nakov and Tiedemann 2012). This is because earlier approaches were based on lexical translation probabilities, which require a

meaningful relationship between characters in the source language and characters in the target language, which only holds true for similar languages. Consider this Norwegian–Swedish parallel sentence taken from Tiedemann (2009):

Norwegian: Så du bø være temmelig redd .
Swedish: Så du bör vara mycket rädd .

Note how strong the character-level relationship between the languages is (one English translation could be 'so you should be pretty scared'). Character-based MT between more distant languages has also been shown to be effective through the use of hierarchical models which compose larger units from characters. This allows translation probabilities to be modelled at coarser granularities but without restricting them from modelling character level translation phenomena when appropriate. This has been explored by Neubig *et al.* (2012b), who extend the Bayesian inversion transduction grammar approach described above (which is used in Chapter 3), as well as in more recent character-based approaches to translation that have been used in *neural machine translation* (NMT) contexts.

The most basic neural model is an *encoder-decoder* framework (Kalchbrenner and Blunsom 2013; Sutskever *et al.* 2014; Cho *et al.* 2014). In this framework an *encoder* is used to transform input one-hot vectors representing tokens on the source side into a hidden representation encapsulating the meaning of the whole sentence. A *decoder* then takes this representation of the source sentence and uses it as the basis for generating a sequence of words in the target language. The encoder and decoder are connected end-to-end for training. A popular extension to this model uses an *attention* mechanism (Bahdanau *et al.* 2014) to weight the relevance of source-side words in determining the output, resolving issues such as sentence length that arise from the basic encoder-decoder approach that uses a single vector representation to capture the entire meaning of the source sentence.

Two appealing properties of neural network models are that 1) they share information between words using richer distributed word representations and 2) the architecture is simplified compared to the traditional statistical MT pipeline. However, this second property comes at the cost of the model being less interpretable. In

recent years the performance of word-level neural MT has exceeded that of phrase based statistical MT, with research in the former largely replacing that of the latter.

Recently, character-based neural MT has been explored. This includes a neural MT approach built on the language model of Kim *et al.* (2016b) which involves composing source-language character-level information into word embeddings via a convolutional encoder with highway layers (Costa-jussà and Fonollosa 2016). However, the output vocabulary remains at the word level, and thus suffers from an imposed limit on the vocabulary size in decoding since the models require a probability distribution over each possible target word. It is appealing to deal with characters on the target side as well in order to overcome this problem, since the character inventory is fixed and vastly smaller than the word inventory. Ling *et al.* (2015) do this, using character-level information both for encoding and decoding in an attentional neural MT model. For encoding, a character-level bidirectional *long short-term memory* (LSTM) network is used to compose each word out of character embeddings. At the decoding stage a hidden target word representation is converted into characters using a forward-LSTM conditioned on the word representation and previously generated characters. Thus the model is capable of interpreting and generating unseen word forms. Such models have demonstrated that multilingual many-to-one models can outperform bilingual specific models (Lee *et al.* 2016). However, the training data requirements of such neural MT models are large.

2.2.3 Data Sparsity in Machine Translation

Language documentation work necessarily happens in a low-resource context. In the field of machine translation, there has been a wide variety of work on models to better cope with limited training data. This includes work directly on MT for low-resource languages (Carl *et al.* 2008; Irvine and Callison-Burch 2013; Mikolov *et al.* 2013b; Östling and Tiedemann 2017), but the issue of data sparsity arises in MT even for large languages, and is one of the reasons varying token granularities have been explored. Data sparsity is caused by a variety of factors, including morphological complexity (inflections, compounding) and phenomena such as numbers, proper

names. Because of the infinite productivity of human languages and a long tail of rare words, all machine translation can be considered a battle against data sparsity. In light of this, the division between low-resource and high-resource languages is not clear cut.

Morphology

For languages with rich morphology, such as Turkish, Finnish and Russian there will be many instances of words that haven't been seen many times during training. Approaches to deal with this data sparsity for such languages typically involve segmentation of words into finer grained units such as morphemes (Lee 2004; Zwarts and Dras 2007; Clifton and Sarkar 2011; Ataman *et al.* 2017; Bastan *et al.* 2017) reducing the size of the vocabulary and taking advantage of regularity in the morphology. A popular recent approach is the use of *byte pair encoding* (BPE) in MT (Sennrich *et al.* 2015), which involves iteratively replacing the most commonly paired characters and groups of characters, creating a smaller vocabulary of frequently seen character n-grams (of variable n). Decomposition of words based on BPE has been used on its own, as well as in combination with character-based models, such as in the model of Chung *et al.* (2016), which performs character level and sub-word level decoding, with source sentences segmented using BPE. To compare segmentation granularity for Chinese–English and English–Chinese translation, Wang *et al.* (2017a) explore character-level, BPE-level and word-level Chinese–English translation using LSTM encoders and decoders (after word segmentation with in-house tools).

Other concepts to address morphology include incorporating source-side linguistic information for more accurate generation of morphology in a richer target language (Chahuneau *et al.* 2013; Durrani *et al.* 2014), paraphrasing to effectively create different translation inputs when translating from a morphologically richer language (Nakov and Ng 2011), and modelling the influence of morphemes on nearby words to improve alignment (Luong and Kan 2010). Bergmanis and Goldwater (2017) address morphological analysis by abstracting over spelling differences between functionally similar morphemes, unlike much previous work that focuses solely on segmentation.

Transfer Learning

Transfer learning is another approach to deal with data sparsity which is broadly applicable (including to speech recognition, as we will discuss in §2.3.4), but in machine translation works by applying information learnt from translating some languages to the task of translating others. These approaches commonly involve sharing neural network parameters between models to harness the similarity of the distributional representations of words and that of their translations, and is useful both in high-resource settings (Johnson *et al.* 2016) and in low-resource settings (Zoph *et al.* 2016; Nguyen and Chiang 2017). Zoph *et al.* (2016) demonstrate that training a NMT model for a high-resource language pair can then be used to substantially improve translation performance from a low-resource language into a high-resource language. This allows embeddings in the common high-resource language to be transferred. Though the neural MT alone underperforms a syntax-based alternative, ensembling such pre-trained low-resource neural MT models allows for improvements over the syntax-based alternative. Nguyen and Chiang (2017) expand on this model to exploit vocabulary overlap between related low-resource languages, enabling transfer from one low-resource Turkic language to another to improve translation into English.

2.2.4 Unsupervised Word Segmentation

In the MT frameworks discussed above, translation typically assumes the use of orthographies that include spaces to delimit words. Sometimes such segmentation is not available and automatic word segmentation is necessary. Word segmentation plays a very important role when performing natural language processing tasks on languages such as Chinese and Japanese which do not include segmentation in the orthography. It is also used in segmentation-based approaches to handling morphology discussed in §2.2.3; and is relevant in this thesis since phonemic representations of speech are unsegmented by default.

Word segmentation for languages whose orthographies do not include them can involve the use of a dictionary or segmented training corpus for supervised training

to help determine candidate points in strings at which to split (Nagata 1997; Sassano 2014; Cai *et al.* 2017). However, ambiguities frequently make this task more difficult than one might naively assume. For example, in Japanese **東京都** can validly be segmented as **東** (east) + **京都** (Kyoto), or more commonly as **東京** (Tokyo) + **都** (prefecture). In Chinese, 3.6% of characters could be segmented such that they serve as a prefix of the word to the right, or a suffix of the word to the left (Ma *et al.* 2014).

Unsupervised word segmentation approaches that do not use dictionaries employ statistical methods to lump together high-frequency clusters of characters (Goldwater *et al.* 2006; Johnson and Goldwater 2009; Mochihashi *et al.* 2009; Elsner *et al.* 2013). For such methods, the predicted segmentation of words will frequently deviate from what are considered words in a canonical dictionary. These unsupervised methods find their use in tasks such as modelling the lexical acquisition of children, and preparing text for downstream tasks such as machine translation. Such a segmentation method is used as a component of an alignment approach in Chapter 3.

For machine translation, there is often no single correct segmentation and it has been argued that the best segmentation is the one that leads to the best machine translation scores (Nguyen *et al.* 2010). Xu *et al.* (2004) learn a dictionary from character–word alignment for Chinese–English MT, which is then used to inform segmentation. Xu *et al.* (2008) and Nguyen *et al.* (2010) both describe Bayesian techniques in which bilingual information informs segmentation useful for machine translation, while Chen and Xu (2015) use a variety of features at the character, phrase and sentence levels in a log-linear model. This concept of jointly segmenting while performing non-MT tasks has been explored in other contexts such as segmentation with part-of-speech (POS) tagging (Sun 2011). Su *et al.* (2016) use lattices of possible tokenizations as input to neural machine translation systems to mitigate issues of error propagation.

2.2.5 Bilingual Lexicon Induction

Machine translation techniques such as translation modelling find use beyond the task of classic machine translation of sentences. Bilingual lexicon induction is a task

that involves drawing semantic correspondences between words in a source language and words in a target language. Beyond its implicit use in MT, bilingual lexicon induction has a long history as lexicons play an important role in multilingual tasks such as cross-lingual information retrieval (Levow *et al.* 2005). Since a key task of documentary linguistics is the creation of lexicons to relate the source language with a larger language, we now discuss computational approaches for this task.

The traditional approach to bilingual lexicon induction has involved statistical inference over word-level information and is essentially a post-processing step after word alignment. Early work includes that of Wu and Xia (1994), who explore lexicon induction between English and Chinese using a variation on IBM model 1 to determine translation probabilities after dictionary-based segmentation. A post-processing step uses significance filtering to remove spurious entries by effectively adapting the filtering threshold based on translation entropy. Melamed (1996) present a technique to clean out spurious entries from translation lexicon, highlighting the issue of *indirect associations*, whereby pairs of unrelated words have statistical properties that resemble those of mutual translation. An example of this might be *what* and a translation of *time* being erroneously learnt from a corpus of travel expressions, owing to common sentences beginning with "What time…" Other word alignment based approaches include that of Caseli *et al.* (2006), which involves lexicon induction in the context of translation rule induction for rule-based MT, and that of Lardilleux *et al.* (2010), who compare word alignment tools for the task of bilingual lexicon induction.

Other methods remove reliance on the scarce resource of parallel corpora by using comparable corpora (Fung and Yee 1998; Koehn and Knight 2002) or monolingual corpora (Haghighi *et al.* 2008). More recently, word embeddings (discussed in §2.2.6) have become the main basis for such approaches. Vulic and Moens (2015) learn bilingual word embeddings (see §2.2.6) by adapting the skip-gram model to predict bilingual contexts of words. This method simply involves concatenating a source document with its target translation, shuffling, and then running the skip-gram model training with a large window size. The cosine similarity between words thus corresponds strongly with the likelihood that they are translations, outperforming state-of-the-art bilingual lexicon induction baselines as well as similar bilingual

word embedding models. Bilingual lexicon induction has also been explored with comparable corpora and sub-word information, with Heyman *et al.* (2017) using character level information and multilingual Wikipedia data as a comparable corpus in order to exploit orthographic similarities between words and related languages. Unlike previous work in the same vein, neural models are used to encode character level representations of words for a downstream classification task which determines whether words are translations of one another.

As well as this recent work incorporating character-level information in lexicon induction, there has previously been some work on lexicon induction from phonemes which is closely related to the explorations of Chapter 3. Stüker and Waibel (2008) and Stüker *et al.* (2009) take a first look at phoneme–word translation modelling, using traditional IBM Models (Brown *et al.* 1993) in order to determine alignments, and applying heuristics to extract dictionaries, while Stahlberg *et al.* (2013) build on the approach of Stahlberg *et al.* (2012), using 30k Bible verses for lexicon extraction. None of the above work considers bilingual lexicon induction from speech directly, though we discuss *spoken term discovery* in §2.3.3.

2.2.6 Cross-Lingual Word Embeddings

In recent years word embeddings and cross-lingual word embeddings (CLWEs) have become popular. Relevant to our interests, CLWEs have the capacity to enable transfer learning in cross-lingual contexts, which is relevant in the low-resource domain (Duong *et al.* 2015; Fang and Cohn 2017; Zoph *et al.* 2016).

Word embeddings are vector representations of words in a common vector space such that similar words are closer to one another. They originally arose as a side-effect of neural language models (Bengio *et al.* 2003; Goodman 2001), discussed in §2.2.7. One advantage of word embeddings is that they can help models cope with sparse data by sharing information among words with similar characteristics. Although count-based distributed vector representations of words, such as latent semantic analysis (LSA) (Landauer and Dumais 1997), have a long history, word embeddings have become more popular since recent approaches show effective use of shallow neural

network architecture to learn them from large quantities of data (Mnih *et al.* 2009; Bengio *et al.* 2009; Collobert and Weston 2008; Mikolov *et al.* 2013a; Mikolov *et al.* 2013c). These prediction-based word embedding models, in contrast to traditional count-based models such as LSA, have led to improvements in many natural language processing tasks through their use in initializing the parameters of neural network models when supervised training data is limited, harnessing information from large amounts of unlabelled data (Frome *et al.* 2013; Zhang *et al.* 2014; Zoph *et al.* 2016; Lau and Baldwin 2016).

The most well-known of these approaches are the continuous bag-of-words models and the skip-gram models, which predict words given contexts and contexts given words respectively (Mikolov *et al.* 2013a). In the original formulation, these contexts are sliding windows of words, but subsequent work has used other structural representations as contexts in similar models, such as using dependency parsers for more syntactically motivated word embeddings (Vulić *et al.* 2016). The success of embeddings has led to a myriad of other popular word embeddings approaches, both count-based and prediction-based (Chen *et al.* 2013; Pennington *et al.* 2014; Shazeer *et al.* 2016; Bhatia *et al.* 2016).

Cross-lingual word embeddings (CLWEs) have also been the subject of significant investigation. These methods embed words across two or more languages in a common vector space, such that words and their translations have similar vector representations. Many methods require parallel corpora or comparable corpora to connect the languages (Klementiev *et al.* 2012; Zou *et al.* 2013; Hermann and Blunsom 2013; Chandar AP *et al.* 2014; Kočiský *et al.* 2014; Coulmance *et al.* 2015; Wang *et al.* 2016), while others use bilingual dictionaries (Mikolov *et al.* 2013b; Xiao and Guo 2014; Faruqui and Dyer 2014; Gouws and Sogaard 2015; Duong *et al.* 2016b; Ammar *et al.* 2016; Fang and Cohn 2017). Other methods use neither parallel data nor bilingual dictionaries, instead learning cross-lingual word embeddings through analysis of the monolingual distribution of words in more than one language (Barone 2016; Conneau *et al.* 2017). Most relevant to this thesis are the methods that use bilingual dictionaries to bridge between monolingual corpora, with the approach of Duong *et al.* (2016b) being used to train the CLWEs used in Chapter 4.

2.2.7 Language Modelling

Before moving to speech recognition, we first address language modelling, which is an important component in both machine translation and speech recognition, and other text generation tasks. Language models (LMs) are a tool that score the fluency of some generated text, associating with it a likelihood of it occurring in the language the LM was trained on.

Statistical language modelling was popularised with the advent of large digital corpora and the rise of corpus statistics around 1990 (see Kneser and Ney (1995) and related papers cited there). Traditional language models rely on the concept of an n-gram, which is a sequence of n words. Corpus statistics on the occurrences of n-grams is used to estimate the probability of each word (in the language that the corpus represents) given that preceding $n-1$ words. This allows our models to incorporate the intuition that given a context "I like to eat" the following word is more likely to be "sushi" rather than "run" or "aeroplane." Typically n is not larger than 5 or so due to data sparsity, since longer sequences are unlikely to be seen more than once, which makes statistics unreliable.

There has been much work on n-gram language models, including smoothing approaches to account for data sparsity of higher order n-grams (or infinite order models such as that of Shareghi *et al.* (2016)). Approaches typically involve heuristic back-off to (or interpolation with) word probabilities conditioned on smaller contexts where more data is available (Chen and Goodman 1999; Goodman 2001), as well as work on fully Bayesian models, such as the hierarchical Pitman-Yor processes of (Teh 2006).

Bengio *et al.* (2003) introduced neural language models (along with the notion of word embeddings discussed in §2.2.6), which play an important role in the language modelling experiments of Chapter 4. Rather than representing words discretely and completely distinctly from one another, words would be embedded into a vector space such that similar words have similar vector representation. Thus language models can have smoother distributions and the probabilities of word generation can harness statistics from semantically similar words. However, in practice count-based methods with advanced smoothing techniques remained state-of-the-art due to the large

data and computational requirements of neural language models and the difficulty in effectively training them.

Neural language modelling has since demonstrated powerful capabilities at the word level (Mikolov *et al.* 2010). Notably, long short-term memory (LSTM) models (Hochreiter and Schmidhuber 1997) have been shown to be effective for modelling long-ranging statistical influences that traditional n-gram or log-linear models are unable to model, such as matching closing parentheses to opening ones (Osband *et al.* 2016; Zaremba *et al.* 2014). De Mulder *et al.* (2015) survey neural networks for language modelling.

Such models have also demonstrated effective modelling at the character level (Martens 2011). Such work has many of the same motivations to that of character level and sub-word MT discussed in §2.2.2 and §2.2.3: words have sub-word structure that can be captured and modelling these directly can help address the data sparsity concerns that arise from word-based models. Kim *et al.* (2016b) compose word embeddings out of character embeddings using a convolutional layer and highway network to create a word-level output distribution (which has been used in MT models. See §2.2.2). Other work has involved combining word and character-level information as input. For example, Verwimp *et al.* (2017) concatenate word embeddings and character embeddings of the word to harness this sub-word information before word-level prediction. Lankinen *et al.* (2016) apply character-level models to the morphologically rich Finnish language, using LSTMs to encode an internal word representation before character-level prediction on the output. They demonstrate correct inflections not present in the training data can be scored better than incorrect ones. The character-based predictive approach of the last paper is promising, since out-of-vocabulary words can be generated. This is important for tasks involving generation of words and is promising for low-resource situations where many words are out-of-vocabulary: consider the implicit language model included in the neural MT framework of Ling *et al.* (2015) discussed in §2.2.2. However, the sub-word information need not be at the character level. In languages such as Korean, syllable and morpheme-level LMs have been used with the argument that character-level information does not effectively capture the context of the word (Yu *et al.* 2017).

Other work has also operated at a morpheme-level granularity (Kirchhoff *et al.* 2006; El-Desoky Mousa *et al.* 2010; Sak *et al.* 2010) or at a granularity mixed between words and sub-word units, with frequent words being included in the lexicon but less frequent words being decomposed (Shaik *et al.* 2011; Mikolov *et al.* 2012). Botha and Blunsom (2014) instead use a compositional approach where word vector representations are comprised of morpheme vectors. These models that harness sub-word information are useful for overcoming the data sparsity problem associated with low-resource language modelling.

In addition to these models trained on text, there has also been work on learning language models from speech recognition phoneme lattices (Neubig *et al.* 2012a). The idea underlying this approach is appealing because it allows for language modelling over word-like units even when a lexicon is not available, suggesting applicability of the concept to low-resource languages. They use a hierarchical Pitman-Yor process to learn a lexicon and language model by capturing all the information in a lattice. Blocked Gibbs sampling is used for inference in a weighted finite-state transducer (WFST) framework, where the WFST is created by composing the phoneme lattice with both a (dynamically evolving) lexicon FST and LM finite-state acceptor. This approach has particular relevance to the translation modelling method proposed in Chapter 5, which can be seen as a bilingual variation of this language model.

In low-resource settings, Gandhe *et al.* (2014) investigate neural network LMs, comparing them with count-based language models, and find that neural network LMs interpolated with count-based methods outperform standard n-gram models even with small quantities of training data, while Hao Fang *et al.* (2015) harness sub-word morphological information in neural network models (among other models), outperforming count-based methods without interpolating probabilities. Kurimo *et al.* (2016) investigate a variety of relevant topics for low-resource language modelling, including language model adaptation and decoding with sub-word units on account of the rich morphology of the Finnish language. Also having been explored for low-resource language modelling is cross-lingual language modelling, with work on interpolation of a sparse language model with one trained on a large amount of translated data (Jensson *et al.* 2008), as well as integration of speech recognition with

translation to harness statistics from target language corpora (Jensson *et al.* 2009; Xu and Fung 2013). Bellegarda (2004) review language model adaptation, and argue that small amounts of in-domain data are often more valuable than large amounts of out-of-domain data, but that adapting background models using in-domain data can be even better.

2.3 Automatic Speech Recognition

Having considered work relevant to modelling the bilingual nature of the data used in this project, we now consider work relevant to modelling speech, since language documentation starts with the collection of speech. In this section we overview work on the traditional speech recognition problem, as well as unsupervised speech modelling and adaptation of speech technology to low-resource domains.

2.3.1 Traditional Speech Recognition

Speech recognition has traditionally been framed as the problem of predicting a sequence of words $\boldsymbol{w}$ given a representation of the acoustic signal $\boldsymbol{x}$. There are many challenges in this problem: determining an effective feature representation of the acoustic signal, accounting for interference such as reverberations, ambient noise, handling variable speaker features such as prosody, disfluencies, pitch and the rate of speech, Lombard reflexes, telephone speech, speaker intoxication, dialectal differences and accents along with addressing general difficulties of language such as an unbounded lexicon (Goldwater *et al.* 2008; Kunze *et al.* 2017; Sriram *et al.* 2017). Early work on speech recognition focused on limited domains, with an extreme example being single digit recognition. At the other end of the spectrum is large vocabulary continuous speech recognition (LVCSR) which is a much harder problem, since words may be spoken from a larger lexicon, covering a wide variety of topics, while phenomena such as coarticulation across word boundaries compound the difficulty.

The problem of speech recognition is typically framed as finding the most likely

word sequence $\boldsymbol{w}$ given acoustic features $\boldsymbol{x}$:

$$\underset{\boldsymbol{w}}{\operatorname{argmax}}\, P(\boldsymbol{w}|\boldsymbol{x}). \tag{2.1}$$

Bayes' rule is used to factorise this probability

$$\underset{\boldsymbol{w}}{\operatorname{argmax}}\, P(\boldsymbol{w}|\boldsymbol{x}) = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{\phi})P(\boldsymbol{\phi}|\boldsymbol{w})P(\boldsymbol{w}) \tag{2.2}$$

where $\boldsymbol{\phi}$ is a sequence of phonemes that may have been expressed in the speech. This factorisation is useful for learning model parameters. $P(\boldsymbol{x}|\boldsymbol{\phi})$ is described by an acoustic model, $P(\boldsymbol{\phi}|\boldsymbol{w})$ by a pronunciation lexicon and $P(\boldsymbol{w})$ by a language model.

The prevailing framework that has existed from the 1990s until today uses hidden Markov models to represent a) the lexicon and language model with transition probabilities between words and the phones in those words (the hidden states) (Lee 1990), and b) observation probabilities being described by an acoustic model. Traditionally, Gaussian mixture models were the most popular method for representing an acoustic signal given the sub-phone state, but in recent years deep neural networks have demonstrated superior performance in light of advances in the availability of data, computational power, and the algorithms used to train the networks (Hinton *et al.* 2012).

The representation of the speech signal, $\boldsymbol{x}$, modelled by the acoustic model is represented by a sequence of vectors. There are a variety of representations based on human perception such as Mel-frequency cepstral coefficients (MFCCs) and perceptual linear predictive (PLP) features (Hermansky 1990).

Mainstream speech recognition focuses on word-level outputs, as opposed to a sequence of phonemes. This is because in most domains we are interested in orthographic transcriptions of speech. An advantage of operating at the word level is that there is a constraint on the phoneme sequences being consistent with possible words, and this helps to reduce the search space. The disadvantage is that the lexicon of the language must be known in advance, along with the pronunciations of each word in the lexicon. In low-resource contexts the lexicon may be sparse and thus speech must be effectively recognised at the phoneme level.

2.3.2 End-to-End Speech Recognition

As mentioned above, deep neural networks have, in most places, superseded the use of Gaussian mixture models for modelling the relationship between the signal and hidden states. There is additional growing work on so-called *end-to-end* speech recognition, which involves direct discriminative training of a model to predict the transcription of an utterance given the acoustic signal using a single neural network architecture. A motivation for this end-to-end training is that the components of the speech recognition system are jointly optimized with the target objective in mind, rather than being trained separately on different objectives, leading to "compartmentalization" that can negatively affect results (Miao and Metze 2017).

Graves *et al.* (2006) introduce the connectionist temporal classification (CTC) loss function, which allows recurrent neural networks (RNNs) to be trained without segmentation of the label sequence, even though there is a significant mismatch in the number of input acoustic feature vectors and output labels. This opened the door to body of work on CTC-based models with underlying RNNs in order to predict phonemes (Graves *et al.* 2013). This set a new state-of-the art for results on the TIMIT dataset, while facilitating simpler formulation of speech recognition models that do not need phoneme segmented training data. Other work extends on this, exploring other underlying neural architectures including convolutional neural networks (CNNs) (Li and Wu 2016; Zhang *et al.* 2016c; Zhang *et al.* 2016b). Neural networks trained end-to-end with the CTC loss function in order to predict phonemic transcriptions play an important role in Chapter 6.

CTC has also been used in the context of orthographic transcription prediction, either by incorporating a lexicon and language model (Miao *et al.* 2015; Zweig *et al.* 2016; Wang *et al.* 2017b) a character-level language model (Zenkel *et al.* 2017), or by discarding the concept of a phone and performing direct character-level prediction (Hannun *et al.* 2014; Graves and Jaitly 2014; Maas *et al.* 2015), word-level prediction (Soltau *et al.* 2016), or prediction of multi-character units (Liu *et al.* 2017b). Audhkhasi *et al.* (2017) perform direct word based prediction using CTC, presenting techniques to mitigate the issue where the number of possible output words requires

more training data. Importantly, results still underperform use of phone-based CTC and 4-gram language models.

Beyond CTC, there has been exploration in various other approaches to end-to-end speech, including neural attention (Chorowski *et al.* 2015; Chan *et al.* 2015; Chorowski and Jaitly 2016; Bahdanau *et al.* 2016; Shan *et al.* 2017), segmental models (Lu *et al.* 2016; Tang *et al.* 2017), multitask combinations of CTC with attention (Kim *et al.* 2016a) and segmental models (Lu *et al.* 2017), and discriminative end-to-end training of lattice free maximum mutual information (LF-MMI) models (Hadian *et al.* 2018). Variations on CTC have also been proposed such as the simplified AutoSegCriterion of Collobert *et al.* (2016), who use the CTC-based objective in conjunction with CNNs as the underlying neural network. A key insight from exploration in this space, corroborated by experiments with various architectures, is that direct character prediction can be effective, bypassing phonemic annotation and allowing for a simpler overall speech recognition architecture. This grapheme-level prediction is useful for low-resource written languages, but is not directly relevant to the scenario we focus on in this thesis.

2.3.3 Unsupervised Speech Modelling

Even for languages where we are fortunate enough to have a small amounts of transcribed speech, the amount of untranscribed data will always be much greater. There has been some work on semi-supervised speech recognition, which in addition to the transcribed data harnesses this untranscribed data for improved speech recognition, and is able to reach par performance with less labelled data through the use of autoencoders on phoneme transcription (Dhaka and Salvi 2016; Tietz *et al.* 2017), as well as LVCSR using transcription hypotheses from initial models for retraining (Thomas *et al.* 2017; Nallasamy *et al.* 2012).

In the unsupervised case, with a total absence of labelled data, there are two main areas of work: unsupervised term discovery, and sub-word unit discovery. Both areas of work are motivated by its applications to low-resource languages where annotated data may be very scarce, as well as its potential in modelling human cognition

and early-age learning. Unsupervised term discovery seeks to discover new words in the acoustic signal. These words can be found directly from speech features, using approaches based on segmental dynamic time warping (Park and Glass 2008; Jansen *et al.* 2010) or acoustic word embeddings (Levin *et al.* 2013; Kamper *et al.* 2017); or from 1-best transcriptions (Godard *et al.* 2016); or alternatively from phoneme lattices (Neubig *et al.* 2012a) (see §2.2.7 for more on this last approach).

Sub-word unit discovery, on the other hand, involves learning a relatively small set of phonetic units that can be composed together to create the words in a language (Varadarajan *et al.* 2008; Kempton and Moore 2014; Liu *et al.* 2017a). For example, Lee and Glass (2012) segment speech and learn sub-word units using Bayesian non-parametric methods in a hidden Markov model (HMM) framework. The approach learns sub-word units that correlate closely with English phones, which suggests it can find appropriate phone-like units in contexts where we might not know appropriate units in advance. This can be motivated in the context of endangered languages, where we might not always be able to assume prior knowledge of a useful phoneme inventory. While such approaches to sub-word unit discovery are promising given the inevitably low amount of transcribed data available, in the most language documentation contexts a linguist will be able to determine the phonetic categories of the language in a reasonably short amount (such a step can be considered a constant cost per language, unlike the manual transcription of recorded speech). However, even when an appropriate phoneme inventory is known in advance, semi-supervised approaches harnessing untranscribed speech on top of existing phonetic knowledge make for an appealing line of research relevant to language documentation.

2.3.4 Low-Resource and Multilingual Speech Recognition

Just as data sparsity was an issue in MT, so too is it in ASR. There has been a variety of work on developing speech recognition systems for low-resource languages, even widely spoken languages for which there are many millions of speakers but for which limited resources exist to train ASR systems, such as pronunciation dictionaries (Le and Besacier 2009). This includes collecting data directly from speakers for train-

ing traditional systems for as-of-yet unexplored languages (Laleye *et al.* 2016). "The biggest cost factor in such a development is the need of training data for the acoustic model" (Grézl *et al.* 2011), therefore much responsibility lies in the approaches to data collection for acoustic model training, such as the *Woefzela tool* (De Vries *et al.* 2014) (see §2.1). Beyond speech data, collection and effective preprocessing of text data from the Internet has been explored for low-resource languages (Le and Besacier 2005; Gauthier *et al.* 2016; Kurimo *et al.* 2016). For a review of ASR for low-resource languages, see Besacier *et al.* (2014).

While much of this language-specific work is useful, promise lies in work in two close (even overlapping) topics: *domain adaptation*, whereby the goal is to adapt existing acoustic models trained on large amounts of data to target domains with little data, and *multilingual acoustic modelling*, whereby an acoustic model is trained on data from more than one language. This distinction has also been referred to as that of "model adaptation" versus "heterogeneous transfer" (Kunze *et al.* 2017). In both cases, the idea is to harness a broader range of information to improve performance in a specific domain, or to generalize better across domains.

Domain adaptation may occur in a monolingual context. This includes work on adapting the acoustic model to the properties of a given speaker, such as accent, pitch and pronunciation variation (Sakti *et al.* 2011; Hofmann *et al.* 2012). However, it also may involve adapting an acoustic model trained on one language to effectively recognise the sounds of other related languages (Imseng *et al.* 2014; Scharenborg *et al.* 2017) and more distant languages (Schultz and Waibel 2001a; Le and Besacier 2005; Stolcke *et al.* 2006; Tóth *et al.* 2008; Plahl *et al.* 2011; Do *et al.* 2014b).

In multilingual acoustic modelling an acoustic model is trained in data from more than one language. The expected advantages this approach has are the same as those for multitask learning more generally: prevention of overfitting, eavesdropping (attempting other related tasks can guide the model to informative features), and data amplification (different noise added to same feature may help) (Heigold *et al.* 2013). There is overlap between these conceptual categories of adaptation and multilingual acoustic modelling. For example, Sam *et al.* (2012) perform unsupervised acoustic model adaptation for non-native speakers using multilingual acoustic models, while

Thomas *et al.* (2012) train a multilingual model on large amounts of Spanish and German data before adaptation to 1 hour of English (Thomas *et al.* (2017) do similar multilingual + domain adaptation). Such approaches are not exclusive to low-resource languages: Bohac *et al.* (2014) use multilingual data to create systems to rapidly adapt to people with different speech impairments.

In training an acoustic model to recognize the sounds in multiple languages, there are two main approaches that can be taken: common phonemes between languages can be merged into a single symbolic representation (say, based on their representation in the international phonetic alphabet) (Köhler 1998; Lin *et al.* 2009; Grézl *et al.* 2011), despite differences in articulation between the languages and allophonic variation within those languages. The other approach is to treat all phonemes in each language as symbolically distinct from all the phonemes in all the other languages (Thomas *et al.* 2012; Heigold *et al.* 2013; Ghoshal *et al.* 2013; Huang *et al.* 2013; Xu *et al.* 2015; Sercu *et al.* 2016). In the latter case, the acoustic models typically share information in lower layers of a neural network, but have separate layers for language-specific prediction, which can also yield improvements. (Vu *et al.* 2014) compare the two approaches with varying results depending on how related the languages are. For non-related languages in a low-resource rapid adaptation setup most relevant to the work in this thesis, they found no merging strategy consistently performed the best. Vu *et al.* (2012) use a combined approach, where a merged universal phoneme inventory between languages is used for multilingual training, before adaptation and extension of the phoneme set to a low-resource language not in the original training set. Multilingual models have also been used in end-to-end frameworks. For example, Toshniwal *et al.* (2017) use the attentional model of Chan *et al.* (2015), consistently outperforming language-specific models.

Despite this large body of work, there currently exists no freely available universal phoneme transcription tool that can be used by linguists in an automated or semi-automated transcription workflow.

The principle behind multilingual acoustic modelling is to generalize better. Even monolingual speech recognition systems typically aim to generalize well to voices unheard in the training data. However, sometimes generalization isn't the goal. In

language documentation settings, there may be just a few speakers whose speech is being recorded and forming the basis of linguistic analysis. While multilingual acoustic modeling may help generalize across speakers, it is questionable how well a multilingual acoustic model can help with speech recognition for such single-speaker contexts as opposed to a model trained on just a small amount of in-domain data.

In addition to automated approaches to transcription, there has been work on crowd-sourcing transcriptions in a low-resource domain from non-native speakers (and in fact speakers with no knowledge of the target language at all). If we view the human transcribers as acoustic models, then this is analogous to cross-lingual acoustic modelling (Jyothi and Hasegawa-Johnson 2015; Liu *et al.* 2016).

2.3.5 Tonal Speech Recognition

Tonal modelling presents a challenge for ASR systems since tonal information is suprasegmental, spanning many frames (Hu *et al.* 2014; Mortensen *et al.* 2016). Most work in tone recognition sits in the context of speech recognition, though there also has been work on tone-only transcription (Bird 1994).

Advances in tone recognition have involved improving approaches to extracting pitch features from the waveform (Huang and Seide 2000; Lei *et al.* 2006). However, ASR systems often do not incorporate pitch information for speech recognition of tonal languages (Metze *et al.* 2013), as spectral information has performed adequately, with methods instead relying on contextual information for tonal disambiguation via the language model (Le and Besacier 2009; Feng *et al.* 2012) and tone modelling in language-dependent setups, for example accounting for language-specific tone sandhi (Lamel *et al.* 2011). However pitch information has been shown to be useful for even non-tonal languages in neural network frameworks, where prosody is traditionally perceived only to be valuable at the sentential level. To this end Metze *et al.* (2013) explore explicit (*tone-tag*) and embedded (*tonal phone*) modelling approaches for tonal languages (a common distinction between approaches is that between *embedded* tonal modelling, where phoneme and tone labels are jointly predicted, and *explicit* tonal modelling, where they are predicted separately (Lee *et al.* 2002)) finding that

pitch information does improve speech recognition for non-tonal languages and that the *tonal phone* model robustly performed better. The work of Ghahremani *et al.* (2014) corroborates this: they propose a method for pitch feature extraction, and find that along with strong improvements for tonal languages when tone is marked in the lexicon, modest improvements for non-tonal languages are found too. Pitch features as input features in tonal prediction is are used in experiments in Chapter 6.

2.4 Machine Translation Meets Speech Recognition

So far we have discussed MT and ASR separately, but there is much overlap. Both are sequence transduction tasks that involve generating text. Issues that affect MT, such as data sparsity and low ratios of annotated to unannotated data, also affect ASR.

In addition to independent research on speech recognition in machine translation there has also been a significant amount of work on problems that involve both MT and speech recognition. The two most popular problems involving the intersection of speech recognition and MT are speech-to-speech translation and computer-aided translation. In this section we discuss work on these problems. Importantly, in addition to these topics we discuss more recent work on training translation models directly from speech, which is of relevance to the translation modelling method proposed in Chapter 5.

2.4.1 Speech-to-Speech Translation

There has been extensive work on combining ASR and statistical machine translation (SMT) systems, with the work largely focused on coupling the systems for the problem of speech translation (Vidal 1997; Matusov *et al.* 2005; Ney 1999; Casacuberta *et al.* 2004). Speech-to-speech translation involves taking a speech signal in a source language and producing speech in a target language. Traditionally this problem has been approached using speech recognition, MT, and speech synthesis systems in cascade, such that the source speech is first transcribed into text before

text-based translation into the target language before finally applying speech synthesis. However, rather than feeding the one-best speech recognition hypothesis into the MT system, popular approaches change the interface between the systems, feeding instead the speech recognition lattice into the MT system. In this way, the extent to which errors in the speech recognition system are propagated to the system is minimised, since more information in the form of alternative hypotheses are captured in the lattice, which is useful when the 1-best transcription is false.

2.4.2 Computer-Aided Translation

There has also been a variety of work on using translation models to improve ASR performance (Vidal *et al.* 2006; Alabau *et al.* 2011), which includes the popular computer-aided translation (CAT) use case pioneered by (Brown *et al.* 1994). This task involves humans working with computers to perform translation. The human translator is presented with source text to translate and is tasked with producing a spoken translation. The computer then recognises this speech and produces a target transcription. This is fundamentally just a speech recognition task in the target language. However, the speech recognition system is given additional information in the form of the source text which can inform transcription hypotheses in the target language. Work on computer-aided translation has typically involved modifying language model probabilities in the ASR system (Rodriguez *et al.* 2012; Pelemans *et al.* 2015; Ng *et al.* 2013) using N-best lists (Paulik *et al.* 2005). Additionally, word lattice based approaches have also been pursued (Khadivi and Ney 2008; Reddy and Rose 2010). The transcription of multiple streams of interpreted speech has also been addressed with the aid of machine translation (Miranda *et al.* 2012a; Miranda *et al.* 2012b). However, in all of these pieces of work, the translation models are trained on substantial external written corpora such as European parliament proceedings or the Canadian Hansards. In a low-resource context, such information will not be available.

2.4.3 Translation Modelling of Speech

A third area of work in which speech recognition meets MT is work on training translation models on representations of speech and corresponding text. Such work has been motivated by speech-to-speech translation, using word-level automatic transcriptions of interpreted speech to train a translation model more appropriate to the spoken nature of the data (Paulik and Waibel 2009; Paulik and Waibel 2010; Paulik and Waibel 2013). Other work involves source-language phonemic representation along with word-level translations in the target language. Stüker and Waibel (2008) and Stüker *et al.* (2009) use 1-best phonemic representations of the source speech for translation modelling and lexicon induction using traditional IBM models of word alignment. The motivation is that such an approach would allow the written form of the language to be bypassed. This could potentially facilitate speech translation of non-written languages (Besacier *et al.* 2006).

A key challenge in this task is learning an effective translation model given such small amounts of erroneous transcriptions and building meaningful correspondences between phonemes and target words, where there is a substantial mismatch in granularity of the token. To address this, Besacier *et al.* (2006) perform unsupervised word segmentation and discovery on the phoneme sequences before translation modelling. As an alternative model for this task, Stahlberg *et al.* (2012) propose Model 3P, an extension to IBM Model 3 that includes additional word length parameters in the generative model, allowing for more effective alignment between phonemes and words, and word segmentation of phonemes. As well as word-to-phoneme alignment, Model 3P has been used for pronunciation dictionary induction (Stahlberg *et al.* 2013; Stahlberg *et al.* 2014b; Stahlberg *et al.* 2015) to help facilitate speech recognition in a low-resource scenario without target language training data (Stahlberg *et al.* 2014a). Jiang *et al.* (2011) investigate translation using a phonetic representation of the input sentence. However, they train a translation model in a phrase-based SMT framework on standard word-based parallel corpora before converting it into phonemic representations to reduce the ill effects of recognition errors at test time for speech-to-speech translation, thus relying on predetermined word segmentation.

Such work is often motivated by the low-resource bilingual data scenario, as is the motivation for the work in this thesis. However, the approaches listed so far are limited in that they assume an error-prone 1-best phoneme transcription of speech. As we saw in §2.4.1, it is well documented that error propagation can be minimized through the use of lattices or n-best lists instead of 1-best transcriptions.

More recently, and contemporaneous to the work of this thesis, there has also been work on directly relating speech to a target translation without first performing transcription at the phoneme level. In this spirit, Duong *et al.* (2016a) use an attentional model for their proposed task of speech-to-text alignment. Their stacked pyramidal LSTM-based attentional encoder facilitates speech-to-text alignment almost as accurately as using gold source phoneme transcriptions. Though their model has the capacity to learn effective representations for phone recognition, translations were not harnessed to improve this transcription, with the tasks of speech-to-text alignment and translation being emphasized. Anastasopoulos *et al.* (2016) present another model for this task, which marry a reparameterization of IBM Model 2 (Dyer *et al.* 2013) with dynamic time warping for speech-to-translation alignment, outperforming the neural model. This model was subsequently extended upon to provide translations of unlabelled segments of speech (Anastasopoulos *et al.* 2017). Their approach outperforms a model for a similar task which instead uses unsupervised term discovery cascaded with MT (Bansal *et al.* 2017b) instead of using the translations to inform the discovering of units, which has been shown to be beneficial (Bansal *et al.* 2017a). This line of work has parallels to standard unsupervised term discovery discussed in §2.3.3, with the key distinction being the availability of translations for model training, which can help inform the discovery. The discovery of these word units in a bilingual framework allows for partial translation based on word spotting.

More recently still there has also now been work on full translation directly from speech using attentional models (Berard *et al.* 2016). Weiss *et al.* (2017) go further to perform source language speech recognition jointly with direct translation, which has the benefit of using phonemic transcriptions where available in order to train the model for automatic transcription. All of the work in the last two paragraphs involves bypassing a phonemic representation, meaning the system is not constrained

by errors propagated from the acoustic model and generalizes to scenarios where the phoneme inventory is not known or an acoustic model cannot be effectively adapted. Yet for language documentation purposes, there still is an incentive to acquire phonemic transcriptions. These two goals are not inimical. As Weiss *et al.* (2017) show, training phonemic transcription jointly with the translation task when transcriptions are available is beneficial for the translation task. Perhaps phonemic transcription can benefit via joint translation model training too.

The areas we have discussed: data acquisition, translation modelling, speech recognition, and their combination, provide much of the context for the work in this thesis. Key work in this thesis sits within this intersectional area of translation modelling of speech (Chapters 3 and 5), drawing on understanding in machine translation and speech recognition. Other work addresses language modelling, relevant to both speech recognition and MT, while still tapping into bilingual information (Chapter 4). Monolingual speech recognition is explored too, addressing the nuances of tone and the language documentation context (Chapter 6).

Chapter 3

Translation Modelling of Phonemes

Large portions of this chapter have appeared in the following paper:

> Adams et al. (2015) Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions, in *Proceedings of the 12th International Workshop on Spoken Language Technologies*, Da Nang, Vietnam. pp. 248–255.

3.1 Introduction

For low-resource languages, speech recordings are the primary source of data. Spoken translations into a major language, along with transcriptions of these translations, are easy to obtain and—in the case of transcriptions—might be derived automatically. Armed with speech recordings in the source language, along with orthographic transcriptions in the target language, we can start to model the correspondences between them. When little bilingual data is available, we have limited information about these correspondences between the spans of speech on the source side, and the words on the target side. The presence of errors multiplies the indeterminacies. To get started, we introduce two assumptions.

The first assumption is that (a) we have ample quantities of phonetic transcriptions and (b) that these transcriptions are error-free. In this chapter we conduct experiments to assess whether sentences of unsegmented phonemes paired with translations can be effectively modelled given these two assumptions. We then relax the

first assumption and scale the available training set down to as little as 1,000 sentences of parallel data. Such a small quantity of data is a reasonable amount for a project in documentary linguistics to procure manually with high transcription accuracy. This serves as our first step towards more realistic data. Aside from helping to reign in complexity, this lets us isolate the translation modelling aspect of this work from confounding speech recognition factors. Subsequent chapters (Chapter 5 onwards) of this thesis will additionally remove the second assumption of error-free transcription.

This chapter proceeds as follows. As a preliminary step, we first investigate modelling the bilingual data in an end-to-end machine translation context that makes both of the above assumptions. This is done in order to assess how the challenge of coping with unsegmented phonemes can be addressed and whether comparable translation quality can be attained in the absence of segmentation. Can unsegmented phonemes and target words be meaningfully aligned? If so, then perhaps information from the larger language can inform how we process the smaller language. We then relax this assumption of ample data, scaling available data down in the context of a bilingual lexicon induction task. Bilingual lexicon induction is a key step in the language documentation workflow. We compare the strengths and weaknesses of a variety of models, including (a) monolingual segmentation followed by alignment, (b) a model that jointly segments and aligns in a hierarchical inversion transduction grammar framework, (c) traditional IBM Model word alignment models as well as (d) an extension to IBM Model 3 proposed by Stahlberg *et al.* (2012).

3.2 Phoneme-Based Machine Translation

Before we investigate bilingual lexicon induction directly, we begin with a preliminary experiment into translation modelling in an end-to-end machine translation context, demonstrating that translation between foreign phoneme sequences and English is feasible in the absence of word segmentation and punctuation. This is accomplished by applying non-parametric Bayesian methods in an inversion transduction grammar (ITG; see Wu (1997)) framework. Though the later portion of this chapter focuses

on bilingual lexicon induction through construction of a translation model, this constitutes just one main component in an end-to-end machine translation pipeline. We perform machine translation here as a preliminary step in order to get an understanding of how effective unsegmented sequences of phonemes can be modelled if ample data is available, and how it compares to word-based machine translation systems.

There is value in machine translation between endangered languages and English and vice versa in its own right. First, it has the potential to open up a world of educational information to speakers of small languages for which manual translation of a wide variety of materials is impossible, though this involves the more challenging task of translation into the endangered language, for which language model quality will be limited by available data. Second, in some instances it could arguably reduce the rate of language death since there will be more of an incentive for speakers of small languages to continue speaking their language if more information is available in it. Third, it would provide a measure of how well a language has been documented. The more comprehensive the data the machine translation system has to work with, the better the quality the translations can be expected to be. Because of this, shortcomings in machine translation quality can highlight gaps in the collected data, guiding the elicitation of more data. Also, machine translation quality indicates a lower bound on what humans could conceivably learn from the data. If a machine can translate to a certain level using the data, then it could be expected that humans can do at least as well (Abney and Bird 2010).

With the extremely low amount of speech data that is the focus of this thesis, we cannot expect effective end-to-end machine translation. There are many barriers to the task of machine translation of endangered languages. Since we must work with speech, we face the problem of determining appropriate phonemic units for the language and sourcing an acoustic model capable of detecting useful phonemes. However, without a strong pre-existing language model or lexicon, automatically transcribing audio is very error-prone, making word or phrase alignment more difficult. Other issues in speech processing and machine translation are also present, such as determining a useful word segmentation. Furthermore, without linguistic data, actually evaluating the machine translation systems on real-world data is problematic, since

it depends on a pre-existing test set.

Experimental assumptions are made to address the scenario where sufficiently large quantities of data have been acquired by means of this new-wave language documentation effort and use of this data has made it possible for acoustic model errors to be resolved to a high degree of accuracy. Error free transcription of English is also assumed. There is hope that the negative effects of these issues can be mitigated to some degree through an increased rate of data collection due to the proliferation of cheap smartphones and other digital technology, however the remainder of the work in this thesis removes these assumptions.

The work demonstrates that translation from parallel sentences of foreign phonemes (without word segmentation) to English text is possible in two substantially different statistical frameworks and can achieve accuracy approaching that of word–word systems.

3.2.1 Alignment Approaches

We compare two approaches for the task of phoneme–word phrase alignment, which include a traditional maximum likelihood based approach, as well as a hierarchical Bayesian grammar.

IBM Word Alignment Models

GIZA++ (Och and Ney 2003) is the baseline that follows the standard statistical machine translation (SMT) pipeline of performing alignment with the IBM Models (Brown *et al.* 1993). This approach to alignment was used in seminal work on phoneme–word alignment (Stüker and Waibel 2008; Stüker *et al.* 2009). The problem with this approach is that it attempts to capture relationships between individual foreign phonemes and English words, which is extremely difficult.

Figure 3.1 shows the correct alignments for word–word based models and phoneme–word based models. This illustration highlights why token-level alignment at finer granularities is more challenging: there are many more ways things can go wrong. Moreover, the vocabulary size on the phoneme side is lower, making unique alignment

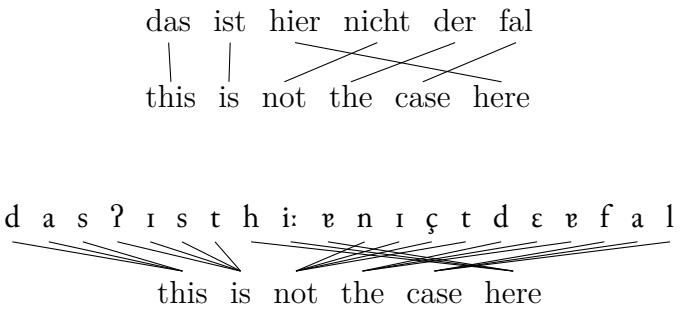


Figure 3.1: Correct alignments for word–word models (top) and phoneme–word models (bottom).

with word-level translations more challenging.

Bayesian Inversion Transduction Grammars

A more promising model for the task of machine translation of phonemic foreign sequences into English is the Bayesian inversion transduction grammar framework of Neubig *et al.* (2011b); Neubig *et al.* (2012b). Alignments are obtained through Bayesian learning of ITGs (Wu 1997) which completely describe the sentence and its translation as a tree of aligned phrases and binary reordering operations and allow for alignment via efficient parsing techniques.

However, in contrast to the preceding work involving Bayesian ITG modelling of many-to-many alignments, where phrases were modelled only at terminal nodes, (Cherry and Lin 2007; Zhang *et al.* 2008; Blunsom *et al.* 2009) the method of Neubig *et al.* (2011b) models phrase-alignments at each node in the ITG using Bayesian non-parametric methods to encourage learning of larger phrase translation units for simpler models, backing off to smaller phrases when appropriate to explain the data. This avoids the issue of only modelling translations of minimal phrases, which otherwise has to be overcome with heuristic phrase extraction methods.

Figure 3.2, illustrates the generative story of a German–English sentence pair, with phrases of different granularities being captured. The Reg and Inv tags illustrate the reordering capacities of the ITG trees, with Reg being a monotone alignment

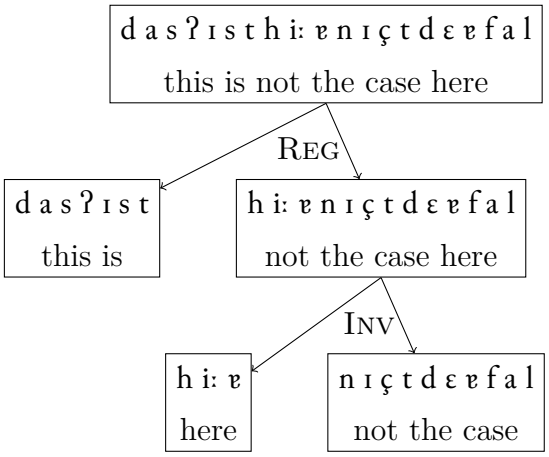


Figure 3.2: An ITG structure learnt by PIALIGN. The phonemes represent the German sentence 'das ist hier nicht der Fall.' Note that PIALIGN forces alignments down to individual tokens, but the leaf nodes presented here represent alignments that were generated as single phrase by the model.

ordering and INV flipping the English side with respect to the foreign phonemes. Allowing flipping at each branch in the tree allows for the vast majority of sentence reorderings to be expressed.

Beyond allowing phrase translations to be captured at varying granularities in a statistical framework free of heuristics, another advantage of this joint learning approach over IBM Model alignment with heuristic phrase extraction (and over UWS GIZA++ discussed in §3.3.1) is that the segmentation on the phoneme side can be informed by the English, which has been shown to be valuable (Xu *et al.* 2008; Chang *et al.* 2008; Nguyen *et al.* 2010).

We use the freely available PIALIGN[1] implementation of Neubig *et al.* (2011b); Neubig *et al.* (2012b). Readers are encouraged to delve into Graham Neubig's thesis (Neubig 2012) for a more in-depth exposition.

[1] `www.phontron.com/pialign/tool`

3.2.2 Experimental Setup

Data

We used 241k sentences of pretokenized German–English training data from the Europarl Corpus (Koehn 2005), with development and test sets taken from the 2005 ACL shared task on machine translation.[2] German data was used since it allowed for scaling up to this substantial amount of data, while permitting easier manual annotation of lexical entries and interpretability of errors. Although German and English are more closely related languages than language pairs encountered in linguistic fieldwork, modelling of the language pair is still complex due to varying word order between the languages and the morphological richness of German relative to English.

For all systems, training data was filtered for sentences where both the source and target side were 100 tokens or less when segmented at the character level. Sentences where the fertility (the ratio of tokens on one side to tokens on the other) was greater than 9 were removed, which constituted a small subset of the training sentences. For the English language model, the full German–English subset of Europarl was used.

For word–word systems, the data was then left as is for training. In the case of phoneme–word translation, the German side of the corpus was converted to a string of phonemes using the MARY text-to-speech system (Schröder and Trouvain 2003), where the phoneme set used is a subset of SAMPA relevant to German.[3] After punctuation and whitespace were removed, the German data was segmented at the phoneme level. Stress markers and syllable boundaries that aren't typical outputs of automatic speech recognition (ASR) systems were removed. The English data remained tokenized at the word level.

Settings

For alignment, GIZA++ (Och and Ney 2003) was used with default settings. In the case of PIALIGN, default settings were used with the exception that the base measure was set to `coocll`, a log-linear interpolation of phrase co-occurrence probabilities

[2] `statmt.org/wpt05/mt-shared-task`

[3] `phon.ucl.ac.uk/home/sampa/german.htm`

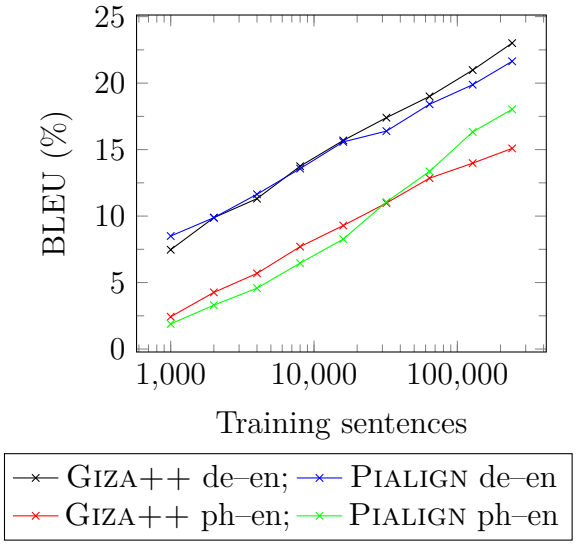


Figure 3.3: BLEU scores comparing phrase tables based on Giza++ and Pialign for phoneme–word and word–word based MT systems translating from German to English.

in each direction, with a discount of 5 (as opposed to IBM Model 1 probabilities as a base measure, which are a poor prior when fine granularity tokens are used (Neubig *et al.* 2012b)).

We used Moses (Koehn *et al.* 2007) for tuning and testing. We performed minimum error-rate tuning (MERT) with cube-pruning for both the Giza++ and Pialign phrase tables. For testing, the decoder was used with default settings.

The English language model was the same for both word–word and phoneme–word systems and was a 5-gram interpolated Kneser-Ney model (Kneser and Ney 1995) learnt using KenLM (Heafield 2011).

3.2.3 Results and Discussion

Figure 3.3 shows the BLEU scores for the word–word and phoneme–word systems as training data is scaled from 1k sentences to ~241k sentences.

For word–word translation, Giza++ and Pialign phrase tables are competitive, with Giza++ performance scaling as an almost perfectly logarithmic function of the

number of training sentences. For phoneme–word translation, performance between the approaches begins to diverge at 128k sentences, when PIALIGN gains a ~2.5 BLEU point advantage.

This advantage of PIALIGN over GIZA++ for phoneme–word translation is expected, due to the aforementioned reason of PIALIGN naturally capturing translation at varying granularities, while GIZA++ produces translation model probabilities just at the token level, requiring subsequent heuristic phrase extraction. In this sense, it is somewhat surprising how well GIZA++ manages to keep pace with PIALIGN on smaller training sets. The use of a common language model in the MT pipeline probably helps to normalize the scores, which are based on word-level English output: as long as there is sufficient signal to produce the output word tokens, the quality of source-side segmentation is not so important. In the next section (§3.3) we evaluate the quality of the entries intrinsically, using human annotators for deeper insight into the quality of the learnt phrases.

While application of GIZA++ allows for phrase-based machine translation, it still depends on combining two sets of one–many word alignments using heuristics to create phrasal alignments. While this is sufficient for word–word translation tasks and char–char translation tasks for similar languages (Vilar *et al.* 2007), the lack of co-occurrence of letters in dissimilar languages makes for poor word alignments. That is, it isn't very helpful to model the probability of an English letter given a German one. It's important to note that this same issue extends to the task of aligning phoneme sequences with words, as lexical translation probabilities that relate phonemes to words have little meaning and make effective modelling impossible.

Most importantly, these results demonstrate that machine translation using sentence-aligned phoneme streams and English is feasible, although the performance significantly underperforms that of traditional frameworks where punctuation and word-level tokenization is present. More importantly for the purposes of this thesis, the results actually suggest that the alignments and phrase tables are meaningful and can be potentially used for other tasks such as bilingual lexicon induction that are less data-onerous than end-to-end MT.

There are a number of ways a phoneme-based machine translation could be cre-

ated. In this section we have evaluated just one setup using PIALIGN and Moses in the interest of demonstrating the feasibility of the approach. Though investigating the efficacy of other phrase alignment tools in the task of translation between foreign phoneme sequences and written text could be done, the point is that the issue of no word segmentation and the fine granularity of phonemes can reasonably be overcome, and in the rest of the thesis we consider other tasks that involve translation modelling but not machine translation.

Experimental results have the limitation that spoken speech takes a different form to written speech converted to phonemes, with spoken speech exhibiting additional complicating features such as coarticulation. Furthermore, while we scale up results to hundreds of thousands of sentences, if such accurate phonemic transcriptions were available, it's likely there would be word segmentation provided by a linguist.

3.3 Phoneme-Based Bilingual Lexicon Induction

In this section we investigate a task more directly relevant to documentary linguistics: creating bilingual lexicons. We consider the task of automatically learning monolingual and bilingual lexical items from unsegmented phonemic transcriptions of bilingual audio where segments of speech in one language are paired with spoken translations in another. In doing so, we remove the first assumption articulated in the introduction of this chapter, which was of an abundance of phonemic transcriptions in parallel with orthographic translations in a larger language. Relaxing this assumption brings us closer to a real-world scenario in its own right, but in addition, it makes the second assumption (that the phoneme transcriptions are correct) more realistic, since the more limited the amount of source-language data, the more reasonable it is that a linguist may manually transcribe it.

Such transcriptions could arise from two scenarios. The first is when future philologists phonetically transcribe speech of a language post-mortem, without native speakers to assist in word segmentation. In such instances lexicon induction would aid in linguistic analysis of the language. The second is by instead employing automatic speech recognition technologies for the same task. In both cases lexicon induction

could aid in bootstrapping ASR systems targeting the language's untranscribed audio. We assume a transcription of the English translation, since English speech can be reliably and cheaply transcribed.

We evaluate existing models that have been used for this purpose in previous work, and report two additional models which demonstrate improvements in lexicon induction. We show that monolingual and bilingual lexical entries can be learnt with high precision from corpora having just 1k–10k sentences. We explain how our results support the application of alignment algorithms to the task of documenting endangered languages.

Previous work on bilingual lexicon induction using sentence-aligned corpora has focused primarily on large corpora of written text (Wu and Xia 1994; Melamed 1996; Caseli *et al.* 2006; Lardilleux *et al.* 2010). Bilingual lexicon induction applied to phonemically transcribed audio, on the other hand, introduces problems including the lack of word segmentation and the small quantities of data. There has been limited work on learning lexicons from phonemic transcriptions. Stüker and Waibel (2008); Stüker *et al.* (2009), mentioned above, take a first look at phoneme–word translation modelling, using traditional IBM Models (Brown *et al.* 1993) in order to determine alignments, and applying heuristics to extract dictionaries. Stahlberg *et al.* (2012) propose Model 3P, which builds upon the generative story of IBM Model 3 by adding additional word length parameters and allowing it to significantly outperform the IBM models (Stahlberg *et al.* 2013; Stahlberg *et al.* 2014a; Stahlberg *et al.* 2014b).

Building on this work, we investigate two models that haven't been considered in this context, and demonstrate that they can outperform the models that have been considered. The first performs unsupervised word segmentation followed by word alignment. The second jointly performs word segmentation and alignment.

Importantly, we evaluate the models on a data set that is significantly smaller than they have been evaluated on previously in this chapter and in other work, containing between just 1k and 10k sentences, corresponding to 13k and 132k words. As a point of comparison, Stahlberg *et al.* (2012) used 123k sentence pairs, Stüker *et al.* (2009) used 146k parallel sentences, Stüker and Waibel (2008) used 155k parallel sentences

and Stahlberg *et al.* (2013) use 30k Bible verses for lexicon extraction. Our training set likely corresponds to roughly 1 to 10 hours of speech (Cieri and Liberman 2006; Bird and Chiang 2012; Bird *et al.* 2014b). These quantities of data are realistic in the context of documentation of endangered languages, though the applicability of these techniques also applies more generally to low-resource languages that have no body of written resources.

In the previous section, we evaluated MT for two different translation models. That extrinsic evaluation yields limited insight to the quality of the lexical entries on the phonemic side, and their implicit segmentation. We now run experiments to intrinsically assess the induced lexicons' precisions at k entries. We do this by applying the alignment models to a German–English corpus, using heuristics (namely, limiting the number of translations of a given word) to extract lexical entries before having them manually annotated.

Results demonstrate that hundreds of bilingual lexical entries can be learnt with good precision, with the additional proposed methods outperforming Model 3P on a data set of 10k sentences. This offers promise of the technique's applicability in a language documentation context. Moreover, the majority of incorrect entries correspond to well segmented, but misaligned, source words.

3.3.1 Translation Models

Our lexicon induction approach uses various phrase alignment techniques to account for the lack of word segmentation and learn phrase translation tables. There are several methods for addressing word segmentation in machine translation (Deng and Byrne 2005; Xu *et al.* 2008; Chang *et al.* 2008; Nguyen *et al.* 2010; Stahlberg *et al.* 2014b), but there has been limited application in a low-resource context. In this section we examine four representative methods to apply to parallel sentences comprised of source phoneme tokens and target words.

Giza++ and Model 3P have been investigated previously for the task of phoneme–word alignment and are evaluated as a point of comparison for the other two methods, UWS Giza++ and Pialign, which we demonstrate are effective for this task.

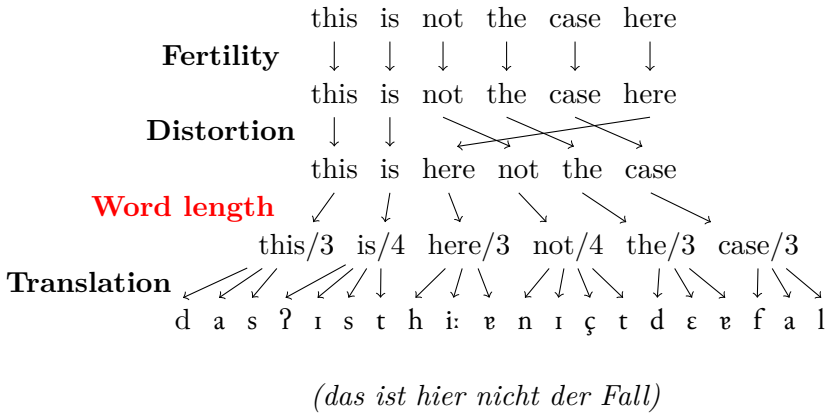


Figure 3.4: The generative model of Model 3P, which extends IBM Model 3 to include additional word length parameters. This allows it to model relationships between source phonemes and target words better than the traditional IBM models.

Model 3P

Model 3P (Stahlberg *et al.* 2012) builds upon the generative model of IBM Model 3 (Brown *et al.* 1993) by adding additional word length parameters (see Figure 3.4), allowing it to outperform traditional IBM models on phoneme–word alignment tasks. After initializing the model with learnt IBM Model parameters, the Pisa implementation of Model 3P[4] uses a genetic algorithm to learn the parameters of the model.

The additional word length parameters, distinct from the fertility parameters (which in a traditional model indicates how many target words a source word is mapped to), allow Model 3P to learn latent word representations that would not be able to be captured in a direct phoneme–word mapping. This allows for better segmentation performance. The key distinction between the fertility parameters and the word length parameters is based on their relationship to the reordering step. Fertility happens before reordering in the generative story. If the word length step also happened before reordering, the phonemes of a word would likely be reordered. Instead, including this step as a separate step from fertility that happens after reordering

[4] `code.google.com/p/pisa`

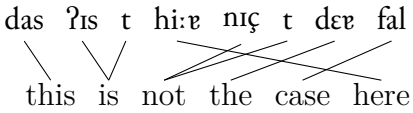


Figure 3.5: Unsupervised segmentation of words followed by alignment, as done in the UWS GIZA++ approach.

preserves the monotonicity of the phonemes representing the word.

UWS Giza++

UWS GIZA++ (unsupervised word segmentation and GIZA++) first performs unsupervised word segmentation using the Bayesian Pitman-Yor language model (Mochihashi *et al.* 2009), as implemented in the tool pgibbs[5] (Neubig 2014). Alignment is then performed between these phoneme sequences and the English words using GIZA++ (see Figure 3.5). This was hypothesized to be more appropriate than GIZA++ alone since it would result in breaking the foreign phoneme sequences into coarser tokens that translate better to English. Note that there is not an expectation that the word segmentation performs well with respect to what is considered a canonical word in the given language. Instead, the key idea is that the segmentation model breaks phonemes into frequently repeating units that capture more meaning than just using individual phonemes. Consider Figure 3.5, where the erroneous segmentation of *ist* and *nicht* nevertheless allows for accurate alignment after monolingual segmentation.

3.3.2 Experimental Setup

Data

To train the translation models we used the German–English parallel corpus from Europarl v7 (Koehn 2005). In order to imitate a phoneme transcription, we converted the German side to a sequence of phonemes (using the SAMPA phoneme alphabet)

[5] `github.com/neubig/pgibbs`

with the MARY text-to-speech system (Schröder and Trouvain 2003). For example, 'dieser' is represented as a sequence of space-separated phonemes, /d iː z ɐ/.

The phonemic output of MARY includes some information that cannot reasonably be detected by an ASR system. In particular, stress markers and syllable boundaries are features output by the system, so we filtered them out. The granularity of tokens on the source side was thus at the phoneme level while English words were used on the target side.

Small quantities of data were used in order to mimic the realities of data collection for endangered languages. We experimented with varying data sizes to evaluate how the best method's performance scales. We used data sets of 1k, 2k, 5k, and 10k parallel sentences (corresponding to between ~13k and ~132k words), a quantity that is vastly smaller than what is typically used in statistical machine translation experiments, but which approaches reasonable size for reliable manual transcription. We limited training sentences to those with fewer than 100 phonemes.

Training

Giza++ was trained using the train-model.perl script included in Moses with default settings, using the `grow-diag-final-and` heuristic for symmetrization/phrase extraction and the msd-bidirectional-fe reordering model.

Model 3P was trained with the Pisa implementation on default settings.

UWS Giza++ was trained by running pgibbs first, and then running Giza++ over the segmented phoneme sequences with default settings. The pgibbs settings were default, with the following exceptions: block sampling was used with a block size of 50, a Pitman-Yor process was used, and 1,000 iterations were run. The final sample output by pgibbs was used as input to Giza++. Giza++ was run in the same way as above, using train-model.perl with heuristics for phrase extraction. It's worth noting that the hyperparameters supplied to pgibbs dictate segmentation granularity. These hyperparameters were not optimized, but were they to change, we would expect the average length of the word units learnt to be different.

Pialign We ran PIALIGN with the base distribution being a log-linear interpolation of phrase co-occurrence probabilities in both directions (with a discount of 5), a beam width of 10^{-6}, a batch length of 40, for 10 iterations. The final sample was used for the purposes of phrase table extraction.

Bilingual Lexicon Extraction

To create bilingual lexicons using the above approaches, entries in the phrase tables were first sorted according to their joint probabilities. We only included entries where the length of the phonemic side was 2 or greater. This heuristic was used since it removed many spurious entries where one foreign phoneme was aligned to an entire word. Additionally, for a given English entry, no more than the top 5 translations were included. A similar filter was applied to prevent more than 5 English translations of a given phoneme sequence. The top 500 entries of each lexicon were then manually annotated.

Annotation

Entries in the lexicon were evaluated by a native German speaker.[6] They were labelled as `Correct`, `Incorrect` or `Ambiguous`. `Correct` entries are those that can readily be found in existing German–English dictionaries. For example, the entry /vɪsən/⇔*know* ('wissen'). `Incorrect` entries are those whose translations are deemed to be clearly incorrect by the annotator. These include entries such as /tsuːʔaɪn/⇔*the* (not a German word) and /bədɪŋʊŋ/⇔*be* ('Bedingung'). In the latter case, note that although the word alignment is incorrect, the phonemes represent a correctly segmented German word, 'Bedingung.'

`Ambiguous` entries are those that are neither strictly correct nor incorrect. These include entries that have boundary errors. For example, /nviːʁ/⇔*we* ('wir') includes an extra /n/ in an otherwise correct entry. Other `Ambiguous` entries are those that, while not found in lexicons, are nonetheless meaningful. These usually highlight

[6] We measured inter-annotator agreement by doubly annotating a sample of 1k entries, using a non-native German speaker, resulting in $\kappa = 0.69$.

interesting linguistic phenomena. For example, /nɪçt/⇔*does not* ('nicht') couldn't be found in Leo,[7] however it captures a meaningful grammatical relationship between the languages. Consider the phrase 'er rennt nicht' and one English translation, 'he does not run,' where this entry makes sense.

3.3.3 Quantitative Evaluation

For the evaluation, precision is favored over recall. Because it is not clear what entries can be reliably learnt from a given bilingual corpus, a measure of recall is difficult to ascertain. In the context of this work it is also more important to determine what can be reliably learnt with high confidence from the small amount of available data.

Precision at k Over Lexical Entries

Figure 3.6 shows the precisions of the bilingual lexicons as the number of entries increases from 1 to 500, using methods trained on 10k sentences. The 'traditional' approach with GIZA++ is the worst performer across the board. This is to be expected as it uses lexical translation probabilities between poorly translated German phonemes and English words as the basis for the extracted phrases. As a point of comparison at the other extreme, using GIZA++ on the gold-standard segmentations of 10k sentences of the original German–English yielded an oracle lexicon with a precision of 0.932 over the top 500 entries.

The other methods are more similar in performance, with the best performing approach being PIALIGN. Though the results are close, the better performance of PIALIGN as compared to the unsupervised word segmentation approach can possibly be attributed to the added information the English side provides in determining useful German segmentation. This contrasts to the unsupervised word segmentation approach which segments using only monolingual German phonemic data. Performance gains over PISA's MODEL 3P can perhaps be attributed to limitations in the generative story of MODEL 3P. Rather than learning explicit phrasal relation-

[7] `www.leo.org`

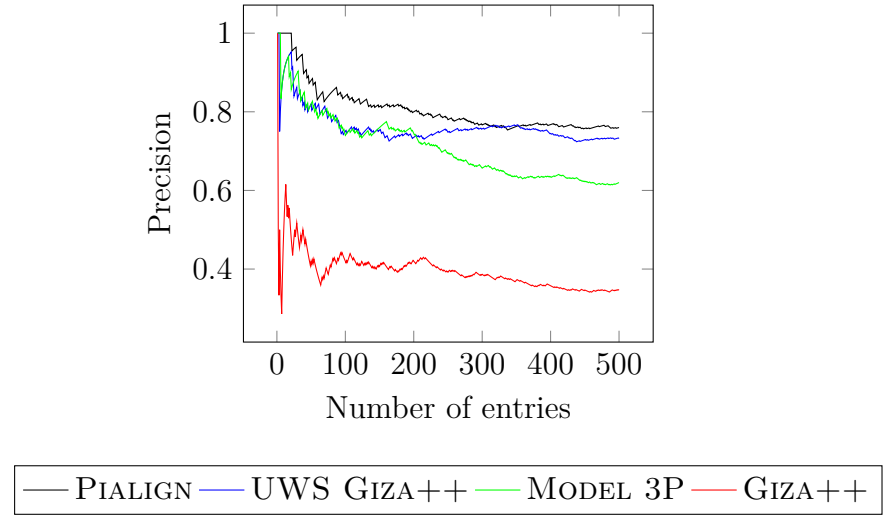


Figure 3.6: Comparison of the lexicon induction methods on the 10k sentence dataset using strict evaluation, where only `Correct` entries improve precision.

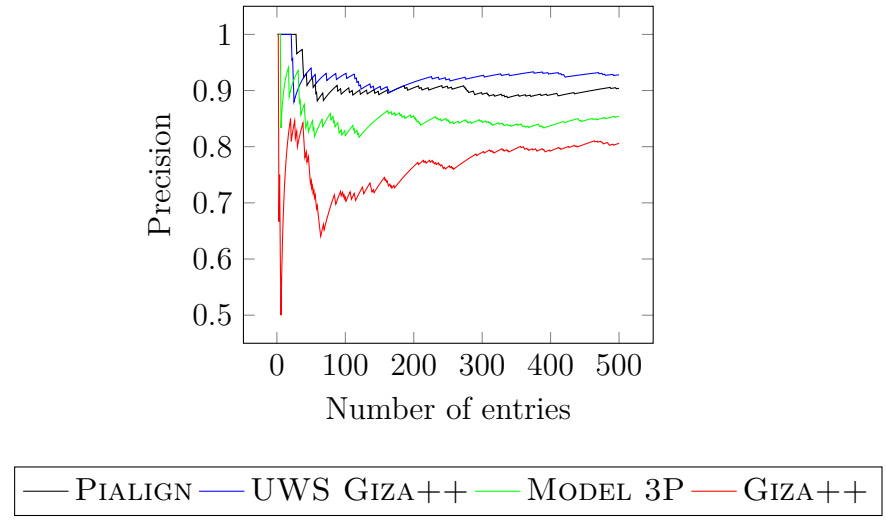


Figure 3.7: Comparison of the lexicon induction methods on the 10k sentence dataset, where `Ambiguous` entries improve precision.

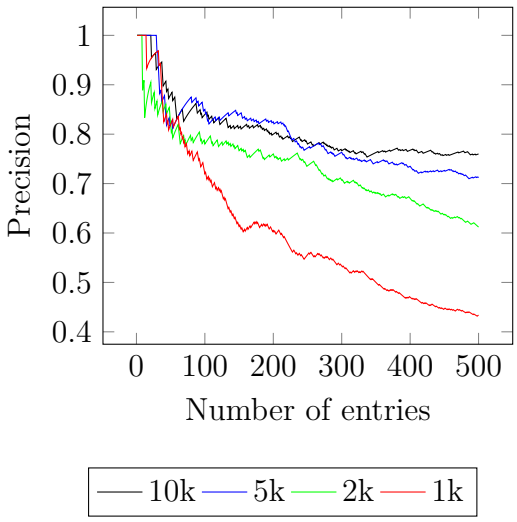


Figure 3.8: Comparison of the precisions of PIALIGN lexicons when learnt from between 1k and 10k sentences.

ships between phoneme groups and words, MODEL 3P conditions the generation of phonemes from latent words and the location within that word. Similar trends in the scores were demonstrated when evaluating precisions that accepted `Ambiguous` entries as also `Correct`. Additionally, in contrast to Model 3P's initialization, which uses the limited phoneme–word alignments of GIZA++, the base distribution PIALIGN draws from additionally uses co-occurrence probabilities of phrases, avoiding this limitation.

Note that when `Ambiguous` entries are considered valid (Figure 3.7), GIZA++ precisions gain the most, and jump up towards 80%. This is due to this approach often including a lot of boundary errors. The phoneme sequence is often incorrectly segmented, leaving promising resulting entries that aren't complete German words. UWS GIZA++ becomes the best performing approach because the issue of under-segmentation propagating to alignment is reduced, since evaluation is softer.

Given that PIALIGN was the best-performing approach on 10k sentences, we additionally evaluated it on smaller data sizes (see Figure 3.8). The fewer sentences of phonemes that are supplied, the more reasonable it is to assume that they can be acquired through reliable manual transcription in a real language preservation scenario. Precision appears to be a logarithmic function of the size of the training data. These

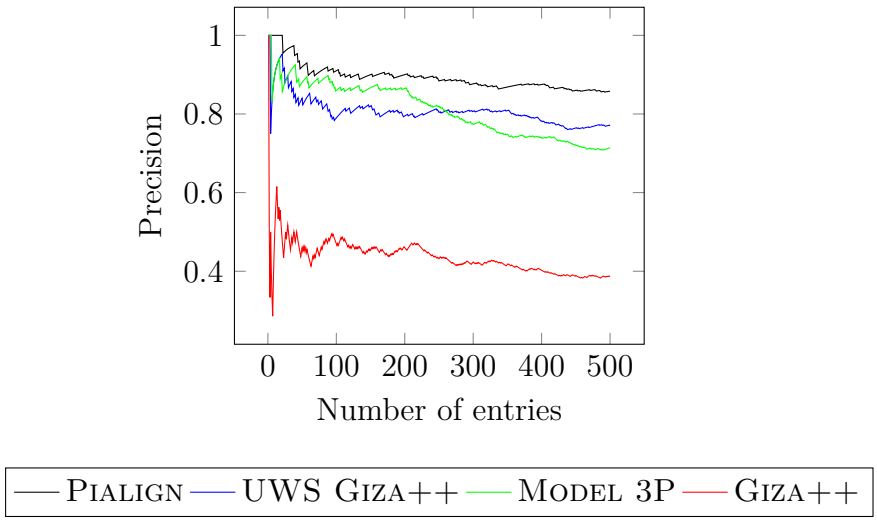


Figure 3.9: Comparison of the monolingual lexicon segmentation precisions for the four lexicon induction methods over the 10k sentence dataset.

results suggest that the first few hundred entries in a lexicon can be acquired with good precision even with very limited data.

Word Segmentation Performance

In addition to evaluating the quality of the bilingual entries, we evaluated the quality of monolingual lexical entries on the phoneme side. This is motivated by the observation that often correct phonemic word units were extracted, but mistranslated. Since monolingual entries are useful in their own right for language documentation purposes (for instance, as a useful starting point for manual correction) and language modelling, we assessed entries that were labelled `Incorrect` or `Ambiguous` to determine whether the phonemic component was segmented correctly at the word boundaries. Note that while it is common to measure token segmentation performance with an F score (as Goldwater *et al.* (2006) do, for example), we are measuring the quality of *types* in a lexicon, and so this should be considered word segmentation in a different sense.

Figure 3.9 shows improved performance of Pialign and Model 3P relative to

		Incorrect		Ambiguous	
Method	Sents	Total %	Acc. %	Total %	Acc. %
GIZA++	10k	19.4	15.5	36.0	0.5
MODEL 3P	10k	14.6	46.6	17.6	13.6
UWS GIZA++	10k	**7.2**	38.9	13.2	3.0
PIALIGN	10k	9.6	**62.5**	**7.8**	**25.6**
PIALIGN	5k	13.4	62.7	9.0	35.6
PIALIGN	2k	16.6	60.2	16.6	18.1
PIALIGN	1k	26.2	52.7	22.2	20.7

Table 3.1: The accuracy of the segmentation of phonemic lexical entries judged `Incorrect` and `Ambiguous`. The *Total %* columns indicate the percentage of entries that were `Incorrect` or `Ambiguous`. The *Acc. %* column indicates the percentage of those `Incorrect` and `Ambiguous` entries that were well segmented, despite not being annotated as `Correct`.

the UWS GIZA++ approach. In the approach of UWS GIZA++, it is impossible to break apart phoneme groups that have been grouped across word boundaries by the monolingual segmentation. However, the other methods aren't constrained by poorly informed early segmentation.

Table 3.1 shows the proportion of the total entries judged to be well segmented but were either `Ambiguous` with boundary errors or `Incorrect`. PIALIGN demonstrates effective inference of lexical items with few boundary errors, outperforming the other methods, regardless of the amount of training data used. This corroborates past research that indicates that word segmentation can be better informed with bilingual data (Xu *et al.* 2008; Chang *et al.* 2008; Nguyen *et al.* 2010).

Although we are evaluating monolingual entries, the entries of UWS GIZA++ are still informed by the alignments with English, as the entries evaluated are the highest probability bilingual lexical entries found. This mitigates the problem of the

Token	Occurrences
ʔ	13,096
ə	8,587
n	8,138
t	6,422
ən	6,300
d	5,929
s	3,226
ɐ	3,136
fl	3,099
diː	2,913

Table 3.2: The most common lexical entries found by the unsupervised word segmentation, without harnessing bilingual information.

effort required to tweak the hyperparameters of the word segmenter to find the right granularity of phoneme clusters. The granularity is instead informed by the English. To appreciate this, consider the most occurring lexical entries of the monolingual supervision *without* being informed by the alignments, as shown in Table 3.2. Of these, the only one that is an actual word is /diː/ (*die*). The rest are common subword units. Note though that /ən/ (*-en*) is a common suffix for infinitive verbs—a particularly useful morpheme.

3.3.4 Qualitative Evaluation

To appreciate the peculiarities and differences of these approaches, we now will consider some general observations made by examining the lexicons of the various approaches, discussing some representative lexical entries and word alignments.

Model 3P seemed generally more susceptible to off-by-one errors at the boundaries of entries. A high confidence—but `Incorrect`—entry that occurred in the lexi-

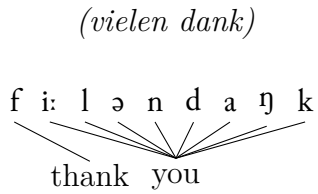


Figure 3.10: The phonemes of *vielen dank* as aligned to *thank you* by MODEL 3P.

con based on MODEL 3P alignments was /iːləndaŋk/⇔*you* (*vielen dank*). The English makes some sense, as 'vielen dank' can be translated as *thank you* or *thank you very much*, although the *thank* component on the English side is missing. Notably, the German side is segmented incorrectly at the phrase boundary, missing the initial phoneme /f/ (it should be /fiːləndaŋk/). It turns out that in sentences containing this German phoneme sequence, the /f/ is often aligned to English *thank* (see Figure 3.10). In the lexicons created by both PIALIGN and UWS GIZA++ this entry was correctly phrase-segmented as /fiːləndaŋk/.

A similar such entry in the MODEL 3P lexicon was /daspaʁlaːmɛn/⇔*parliament*, where the source side is missing the final /t/. In the lexicon constructed using PIALIGN, such boundary mistakes were scarce. The equivalent entry constructed by PIALIGN was /daspaʁlaːmɛnt/⇔*parliament* ('das Parlament'). Note that this entry was not considered strictly `Correct` nor correctly segmented, as it is comprised of two words, with the German article being included (note that the article is optional in English). However in this case, as in almost all others, PIALIGN still segments correctly at the boundaries of multi-word units (as distinct from correctly segmented individual words). The only instance of an entry annotated as `Incorrect` in the top 500 entries of the PIALIGN lexicon where the phoneme side was also incorrectly segmented was /tvoʁdən/⇔*been*, where there is a spurious /t/ prefixing the phonemic representation of *worden*. Investigating the alignments highlights the cause of this entry. Phoneme sequences such as /ʊntɐʃtʏtstvɔʁdən/ (*unterstützt worden*) and /ʔɛʁʁaɪçtvɔʁdən/ (*erreicht worden*) include verbs that often appear inflected with different suffixes elsewhere, but end with a /t/ when occurring before /vɔʁdən/ (*unterstützen* and *erreichen* respectively, with the suffix *-en*). High correlation of /vɔʁdən/ (*worden*) and the suffix

/t/ likely caused this entry.

The lexicon constructed using MODEL 3P demonstrated an apparent bias to shorter units. In that lexicon, the above entry was segmented correctly as /vɔʁdən/. On the other hand, PIALIGN tended to lean towards longer multi-word units as a result of the model's capacity to capture phrases at coarser granularities when it is useful to do so.

/poːɛteːʁʁɪŋ/⇔*Mr Poettering* was present in the PIALIGN lexicon, where the title is missing on the source side. This can be attributed to varying morphology of the title, which takes the form of both 'Herr' and 'Herrn' depending on context. However, since the English side consistently takes the form of 'Mr Poettering', evidence is built up primarily to relate both the title and name on the English side to only the name on the phoneme side.

UWS GIZA++ yielded high confidence, yet erroneous, entries, such as /tʔ/⇔*is*, /nʔ/⇔*to*, /nʔ/⇔*of* that didn't occur in the other lexicons. This is likely a result of the pipelined nature of the approach, where monolingual segmentation is first performed before alignment. The German components to these entries represent frequently occurring phonemic sequences since many words end with /t/ or /n/ and many start with a glottal stop, /ʔ/, before some vowel. The English side represent function words that are so commonly occurring that the coincidental co-occurrence of these phonemes and English words lead them to be learnt by UWS GIZA++ but not by PIALIGN or MODEL 3P. Entries such as this also partly explain why UWS GIZA++ failed to perform as well as MODEL 3P in segmenting lexical entries, despite outperforming it in bilingual precision. The other likely reason is that chunks that cross word boundaries learnt during the monolingual segmentation phase propagate into the translation modelling.

3.4 Discussion

We found that meaningful bilingual relationships can be established for the purposes of machine translation and bilingual lexicon induction despite a mismatch in granularity between the source and target sides. The best performing models are

those that allow flexibility in what granularity is used for modelling (such as PIALIGN), overcame the mismatch by grouping phonemes together into word-like units (UWS GIZA++), or added additional parameters to explain the production of the fine-grained units from the coarse (MODEL 3P).

In the machine translation results of §3.2 GIZA++ did not underperform PIALIGN on low amounts of data. However, in this manual evaluation of the quality of the lexical entries, PIALIGN is substantially better. This shows that comparably accurate segmentation of phrase table entries is not required for comparable machine translation performance with ~10k parallel sentences of training data when measuring the BLEU score of target word-level translations.

3.4.1 Evaluation Issues

Evaluation of the bilingual lexical entries highlighted some issues involved in intrinsically measuring bilingual lexicons. Firstly, the evaluation using *precision*, whereby entries are treated as correct or incorrect, supposes a dichotomy that isn't actually the case. While we can consider lexical entries to be correct if they occur in an established bilingual dictionary, there are many relations between phonemes on the source side and words on the target side that are useful for models to infer based on downstream tasks such as machine translation (such as/nɪçt/⇔*does not* ('nicht')). Such a bilingual lexical item would be useful for machine translation systems as there are many sentences in which this is an appropriate entry in the context of the downstream task (eg. *Er rennt nicht* ("He does not run")).

Word segmentation is another dimension in which learnt lexical entries may be ambiguous. What constitutes a word is ambiguous, particularly in languages that have no standardized orthography. Aside from models committing off-by-one errors frequently, there are also multi-word units that are meaningful and have practical use, despite not aligning in a strictly correct sense to a correct translation. The challenge of assessing correct segmentation and alignment is a burden on the annotator. Though the `Ambiguous` label was introduced, the issue is not solved. If a word segmentation is off-by-one is it `Ambiguous`? What if it's off-by-two?

This highlights an important point about word and phrase alignment algorithms. Such algorithms are rich methods for finding frequent co-ocurrences in parallel corpora (though are to a large extent capable of 'explaining away' non-translation co-ocurrences that frequently co-occur). Incidentally words and their translations frequently co-occur, which is convenient for translation modelling. However, many issues that arise in translation stem from such *indirect associations* (Melamed 1996).

3.4.2 Reconsidering the Value of Bilingual Lexicon Induction

Since downstream extrinsic tasks such as machine translation and speech recognition can benefit from entries in the phrase table that are incorrect when evaluated intrinsically in a bilingual lexicon induction context, it is worth stepping back to consider how bilingual lexicon induction is actually useful to a linguist documenting languages, and how automatically learnt translation models may help that process.

Bilingual lexicon induction is a fundamental task in language documentation. In contrast to phonetic or phonemic transcription, which considers monolingual acoustic properties of the language such as allophonic variations, bilingual lexicon induction relates words in the source language to words in a larger language. This is key for preserving understanding of meaning, since most audio has no grounding in the form of images with which to ascertain meaning from. However, the linguist's creation of dictionaries is much more than the process of collecting tokens and relating them to larger "contact" languages. It involves exploring different senses of the words (perhaps coming up with example sentences), how they might be inflected, and what part of speech they take. They're carefully crafted based on experience with the language. Computers cannot fill the linguist's role here. What computers can do is provide linguists with a large list of statistically sound, unbiased bilingual relationships which can inform their judgements, speeding up the process by making the computer the 'harmless drudge' (Johnson 1755) helping the linguist to better focus on more interesting things.

In the course of transcribing the data that could be fed to models described in this chapter, linguists may have actively been creating a lexicon, since it is one of the

fundamental starting points of language documentation. Indeed a lexicon of modest size may exist before transcription has even begun, or arise during transcription and glossing.

Given that downstream tasks might benefit from entries that would be considered incorrect by the measures used in §3.3, I argue that such simple intrinsic measures of bilingual lexicon quality, while generally informative in the case of large differences (such as those between Bayesian ITG methods and IBM word models in the context of phones) are not particularly useful in the language documentation context. The important insights into the properties of each model came from the qualitative assessment of the nature of the entries each model produced. Providing linguists with such models may be useful for them to ground their understanding of the language based on statistics and help them see patterns they may have overlooked while highlighting differences in the languages. But the choice between such models should not be based entirely on narrow differences in the precision of models.

Therefore for the remaining investigations in the rest of this thesis, we do not measure bilingual lexicon quality intrinsically, but rather consider it as some piece of a larger body of information the models learn or can harness in some other prediction task. In Chapter 5, learning bilingual relations is an integral part of disambiguating the speech signal. In the next chapter (Chapter 4), a presupposed bilingual lexicon containing just word-to-word pairings is used in order to improve language modelling in contexts where there is limited data.

The assumption of error-free phonemic transcriptions limits the insight from the experimentation in this current chapter, and will be addressed Chapters 5 and 6.

Chapter 4

Cross-Lingual Low-Resource Language Modelling

Large parts of this chapter have appeared in:

> Adams et al. (2017) Cross-lingual word embeddings for low-resource language modeling, in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Valencia, Spain.

4.1 Introduction

In Chapter 3 we explored bilingual lexicon induction using small quantities of accurate phonemic transcriptions paired with orthographic translations. However, such data is rarely—perhaps never—available in the absence of other information in the language. Even in the Aikuma-style data collection scenario, typically a linguist documenting the language will have constructed a basic lexicon, and the lexicon will often contain more information than simple bilingual token level mappings, with things such as part-of-speech tags and example sentences. Furthermore, in the process of phonemically transcribing speech the linguist will determine candidate word segmentations. It is phonemic transcription that continues to remain a significant bottleneck for the traditional workflow. In Chapter 5 we investigate methods to directly do this, but in this chapter we explore the use of bilingual lexicons to improve

statistical language models, an important component of speech recognition systems, as well as other tools that involve text generation in the language, such as machine translation systems.

In creating language models for endangered languages, the available textual data is limited to phonemic transcriptions prepared by linguists, since most of the world's languages are not actively written (Bird 2011). Since phonemic transcription is time-consuming, such data is scarce. This makes language modelling, which is a key tool for facilitating speech recognition of these languages, a difficult challenge. One of the touted advantages of neural network language models (NNLMs) is their ability to model sparse data (Bengio *et al.* 2003; Gandhe *et al.* 2014). However, despite the success of NNLMs on large datasets (Mikolov *et al.* 2010; Martens 2011; Osband *et al.* 2016), it remains unclear whether their advantages transfer to scenarios with extremely limited amounts of data, though Gandhe *et al.* (2014) find interpolation of neural network language models with count-based methods to outperform standard n-gram models in low-resource settings and Hao Fang *et al.* (2015) outperform count-based methods without interpolation by using recurrent neural network language models, also in low-resource settings.

Appropriate initialization of parameters in neural network models has been shown to be beneficial across a wide variety of domains, including speech recognition, where unsupervised pre-training of deep belief networks was instrumental in attaining break-through performance (Hinton *et al.* 2012). Neural network approaches to a range of natural language processing (NLP) problems have also been aided by initialization with word embeddings trained on large amounts of unannotated text (Frome *et al.* 2013; Zhang *et al.* 2014; Lau and Baldwin 2016) or on other tasks (Collobert and Weston 2008; Zoph *et al.* 2016). However, in the case of underdocumented languages we do not have the luxury of large quantities of this unannotated text.

As a remedy to this problem we focus on cross-lingual word embeddings (CLWEs), which are learnt using information from multiple languages. Recent advances in CLWEs have shown that high quality embeddings can be learnt even in the absence of bilingual corpora by harnessing bilingual lexicons (Gouws and Sogaard 2015; Duong *et al.* 2016b). This is useful as some threatened and endangered languages have

been subject to significant linguistic investigation, leading to the creation of high-quality lexicons, despite a dearth of transcriptions. For example, the training of a quality speech recognition system for Yongning Na, a Sino-Tibetan language spoken by approximately 40k people, is hindered by this lack of data (Do *et al.* 2014a) despite significant linguistic investigation of the language (Michaud 2008; Lidz 2010) and the availability of an online multilingual dictionary (Michaud 2016).

In this chapter we address two questions. First, is the quality of CLWEs dependent on having large amounts of data in two languages (symmetrical quantities in two languages), or can large amounts of data in a single *target* language inform embeddings trained with little *source* language data (asymmetric quantities)?[1] Second, can such CLWEs improve language modelling in low-resource contexts by initializing the parameters of an NNLM?

To answer the first question, we scale down the available monolingual data of the source language to as few as 1k sentences, while maintaining a large target language dataset. We assess intrinsic embedding quality by considering correlation with human judgment on the WordSim353 test set (Finkelstein *et al.* 2002). For training CLWEs in this chapter, we build on the work of Duong *et al.* (2016b). Their method harnesses monolingual corpora in two languages along with a bilingual lexicon to connect the languages and represent the words in a common vector space. The model builds on the continuous bag-of-words (CBOW) model (Mikolov *et al.* 2013a) which learns embeddings by predicting words given their contexts. The key difference is that the model also tries to predict a source language translation of a target language word centered in a target language context. Since dictionaries tend to include a number of translations for words, an expectation-maximization style training algorithm is used in order to best select translations given the context. This process thus allows for polysemy to be addressed which is desirable given the polysemous nature of bilingual dictionaries. In our work, we remove the assumption that significant monolingual corpora are available on both sides, instead investigating the resilience of such

[1]Note that we have used a different nomenclature to our paper, where *source* was used to denote the large language because it was the source of distributional information in the transfer learning framework. However, to maintain consistency with the rest of this thesis, where *source* is the low-resource language, we instead use *target* to refer to the large language.

approaches in the asymmetric case when one side has scarce monolingual data.

To answer the second question, we then perform language modelling experiments where we initialize the parameters of long short-term memory (LSTM) language models using such CLWEs for low-resource language model training across a variety of language pairs.

Results indicate that CLWEs remain resilient when source language training data is drastically reduced in a simulated low-resource environment (§4.2), and that initializing the embedding layer of an NNLM with these CLWEs consistently leads to better performance of the language model (§4.3). In light of these results, we explore the method's application to Na, an actual low-resource language with manually created lexicons and transcribed data (§4.4). We present a discussion of the negative results found, which highlights challenges and future opportunities.

4.2 Resilience of Cross-Lingual Word Embeddings

Previous work using CLWEs assumes a similar amount of training data for each available language, often in the form of parallel corpora. Recent work has shown that monolingual corpora of two different languages can be tied together with bilingual dictionaries in order to learn embeddings for words in both languages in a common vector space (Gouws and Sogaard 2015; Duong *et al.* 2016b). In this section we remove the assumption of the availability of large monolingual corpora in the source and target languages, and report an experiment on the resilience of such CLWEs when data is scarce in the source language but plentiful in a target language.

The idea underpinning word embeddings is the distributional hypothesis of Harris (1954): that words with similar meaning appear in similar contexts. Prior work has demonstrated the efficacy of cross-lingual word embeddings, suggesting that the distributional hypothesis holds across languages. That is, that words and their translations appear in similar semantic contexts. By greatly reducing the data on the source side, we put this cross-lingual distributional hypothesis to a tougher test of how well CLWEs perform as a vehicle for transferring information from a resource-rich language to a low-resource language.

4.2.1 Experimental Setup

Word embedding quality is commonly assessed by evaluating the correlation of the cosine similarity of the embeddings with human judgements of word similarity. Here we follow the same evaluation procedure, except where we simulate a low-resource language by reducing the availability of source English monolingual text while preserving a large quantity of target language text from other languages. This allows us to evaluate the CLWEs intrinsically using the WordSim353 task (Finkelstein *et al.* 2002), which compares word embedding similarity judgements with those of human annotators, before progressing to downstream language modelling where we additionally consider other source languages.

We trained a variety of embeddings on English Wikipedia data of between 1k and 128k sentences from the training data of Al-Rfou *et al.* (2013). In terms of transcribed speech data, this roughly equates to between 1 and 128 hours of speech transcribed orthographically, with word segmentation. For the training data, we randomly chose sentences that include words in the WordSim353 task proportionally to their frequency in the set. As monolingual baselines, we use the skip-gram (SG) and continuous bag of words (CBOW) methods of Mikolov *et al.* (2013a) as implemented in the Gensim package (Řehůřek and Sojka 2010). We additionally used off-the-shelf CBOW Google News Corpus embeddings with 300 dimensions, trained on 100 billion words.

The CLWEs were trained using the method of Duong *et al.* (2016b) since their method addresses polysemy. The same 1k-128k sentence English Wikipedia data was used but with an additional 5 million sentences of Wikipedia data in a target language. The target languages include Japanese, German, Russian, Finnish, and Spanish, which represent languages of varying similarity with English, some with significant morphological and syntactic differences. To relate the languages, we used the PanLex lexicon (Kamholz *et al.* 2014). Following Duong *et al.* (2016b), we used the default window size of 48 so that the whole sentence's context is almost always taken into account. This mitigates the effect of word re-ordering between languages. We trained with an embedding dimension of 200 for all data sizes as this larger

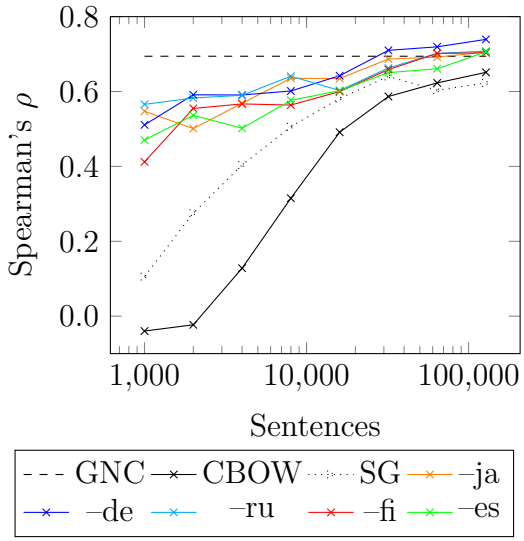


Figure 4.1: Performance of different embeddings on the WordSim353 task with different amounts of training data. *GNC* are the Google News Corpus embeddings, which are constant (having been trained on a 100 Billion word corpus). *CBOW* and *SG* are the monolingual word2vec embeddings. The other, coloured, lines are all cross-lingual word embeddings learnt to harness the information from 5m sentences from one of various source languages (Japanese, German, Russian, Finnish, Spanish).

dimension turned out to be helpful in capturing information from the target side.[2]

4.2.2 Results

Figure 4.1 shows correlations with human judgment in the WordSim353 task. The x-axis represents the number of English training sentences. Coloured lines represent CLWEs trained on different languages: Japanese, German, Spanish, Russian and Finnish.[3]

With around 128k sentences of training data, most methods perform quite well, with German being the best performing. Interestingly the CLWE methods all out-

[2]Hyperparameters for both mono and cross-lingual word embeddings: iters=15, negative=25, size=200, window=48, otherwise default. Smaller window sizes led to similar results for monolingual methods.

[3]We also tried Italian, Dutch, Greek and Serbian, yielding similar results but omitted for presentation.

perform GNC which was trained on a far larger corpus of 100 billion words. With only 1k sentences of source training data, all the CLWEs have a correlation around 0.5, with the exception of Finnish. No consistent benefit was gained by using target languages for which translation with English is simpler. For example, the use of Spanish as a target language often under-performed Russian and Japanese as a target language, as well as the morphologically-rich Finnish, despite English–Spanish being recognized as an "easier" language pair.

Notably, all the CLWEs perform far better than their monolingual counterparts on small amounts of data. This resilience of the source English word embeddings suggests that CLWEs can serve as a method of transferring semantic information from resource-rich languages to low-resource languages, even when the languages are quite different. However, the WordSim353 task is a constrained environment, so in the next section we turn to language modelling, a natural language processing task of much practical importance for low-resource languages.

4.3 Pre-training Language Models

Language models are an important tool with particular application to machine translation and speech recognition. For low-resource languages and unwritten languages, language model quality is poor, since they typically rely on large quantities of data. In this section, we assess the performance of language models on varying quantities of data, across a number of different source–target language pairs. In particular, we use CLWEs to initialize the first layer in an LSTM recurrent neural network language model and assess how this affects language model performance. This is an interesting task for reasons more than just the practical advantage of having better language models for low-resource languages. Language modelling is a syntax-oriented task, yet syntax varies greatly between the languages we evaluate on. This experiment thus yields some additional information about how effectively bilingual information can be used for language modelling.

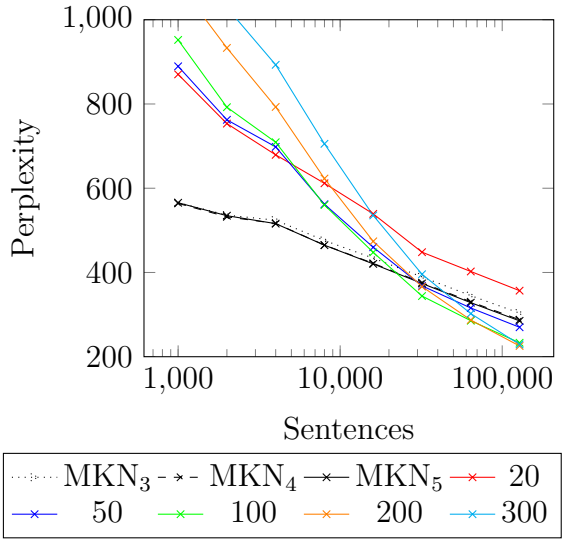


Figure 4.2: Perplexity of language models on the English validation set. Numbers in the legend indicate LSTM language models with different hidden layer sizes, as opposed to Modified Kneser-Ney language models of order 3, 4 and 5.

4.3.1 Experimental Setup

We experiment with a similar data setup as in §4.2. However, source training sentences are not constrained to include words observed in the WordSim353 set, and are random sentences from the aforementioned 5 million sentence corpus. For each language, the validation and test sets consist of 3k randomly selected sentences. The large vocabulary of Wikipedia and the small amounts of training data make this a particularly challenging language modelling task. In our experiments we use a vocabulary of the 10k most frequently occurring words in the corpus, replacing less frequent words with a special rare-word token.

For our NNLMs, we use the LSTM language model of Zaremba *et al.* (2014). As a count-based baseline, we use Modified Kneser-Ney (MKN) (Kneser and Ney 1995; Chen and Goodman 1999) as implemented in KenLM (Heafield 2011). Figure 4.2 presents some results of tuning the dimensions of the hidden layer in the LSTM with respect to perplexity on the validation set,[4] as well as tuning the order of n-grams

[4]We used 1 hidden layer but otherwise the same as the *SmallConfig* of `models/rnn/ptb/ptb_word_lm.py` available in Tensorflow.

used by the MKN language model. A dimension of 100 yielded a good compromise between the smaller and larger training data sizes, while an order 5 MKN model performed slightly better than its lower-order counterparts.[5]

MKN strongly outperforms the LSTM on low quantities of data, with the LSTM language model not reaching parity until between 16k and 32k sentences of data. This is consistent with the results of Chen *et al.* (2015) and Neubig and Dyer (2016) that show that n-gram models are typically better for rare words, and here our vocabulary is large but the number of training sentences is small since the data consist of random Wikipedia sentences. However, the findings from these papers, corroborated further by our findings, are inconsistent with the belief that NNLMs have the ability to cope well with sparse data conditions by using smooth distributions that arise from using dense vector representations of words (Bengio *et al.* 2003).

4.3.2 English Results

With the parameters tuned on the English validation set as above, we evaluated the LSTM language model when the embedding layer is initialized with various monolingual and cross-lingual word embeddings. Figure 4.3 compares the performance of a number of language models on the test set. In every case where pre-trained embeddings were used, the embedding layer was held fixed during training. However, we observed similar results when allowing them to deviate from their initial state. For the CLWEs, the same language set was used as in §4.2. The curves for the target languages (Dutch, Greek, Finnish, and Japanese) are remarkably similar, as were those for the languages omitted from the figure (German, Russian, Serbian, Italian, and Spanish). This suggests that the English source embeddings are capturing similar information from each of the languages, information likely to be more semantic than syntactic, given the syntactic differences between the languages.

We compare these language models pre-trained with CLWEs with pre-training using other embeddings. Pre-training with the Google News Corpus (GNC) embed-

[5]Note that all perplexities in this paper include out-of-vocabulary words, of which there are many. In the model, words not found in the dictionary were given uniform random embeddings between -0.1 and 0.1.

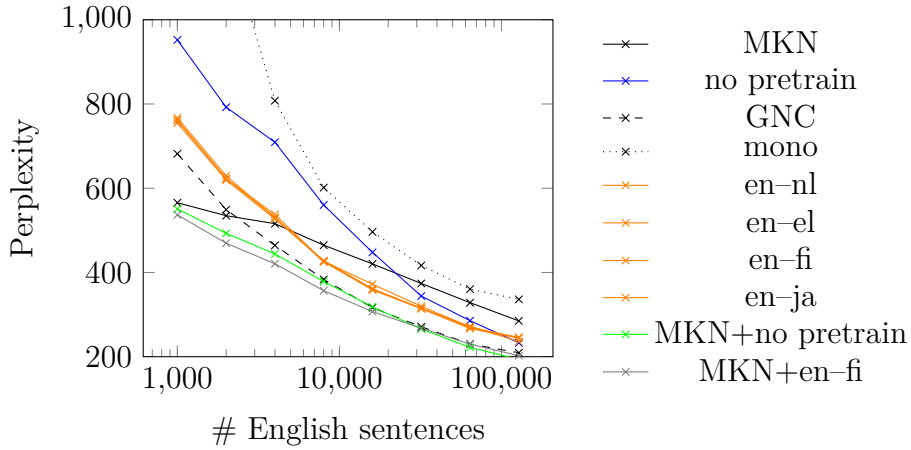


Figure 4.3: Perplexity of English LSTM language models when pre-trained with word embeddings trained on the same English data as the language model. *MKN* is an order 5 Modified Kneser-Ney baseline. *no pretrain* is a neural network language model with no pre-trained embeddings. *mono* is pre-trained with monolingual word2vec embeddings. *GNC* is pre-trained with Google News Corpus embeddings of dimension 300. The rest are pre-trained with CLWEs using information transfer from 5m sentences of Dutch, Greek, Finnish and Japanese respectively. *MKN+no pretrain* and *MKN+en–fi* are results gathered by Adam Makarucha as a point of comparison, which interpolate the probabilities of relevant preceding models.

dings of the method of Mikolov *et al.* (2013c) unsurprisingly performs the best, due to the large amount of English data not available to the other methods, making it an oracle. Monolingual pre-training of word embeddings on the same English data (*mono*) used by the CLWEs yields poorer performance because the embedding layer is held fixed, unlike in the case of *LSTM*.

On small amounts of data the language models initialized with pre-trained CLWEs are significantly better than their counterparts without pre-training, reaching par performance with *MKN* at somewhere just past 4k sentences of training data. In contrast, it takes more than 16k sentences of training data before the plain LSTM language model began to outperform *MKN*. The out-performance of LSTMs by MKN with the lowest amounts of training data motivated interpolation of MKN proba-

	8k sentences				16k sentences			
Lang	MKN	LSTM	CLWE	Interp.	MKN	LSTM	CLWE	Interp.
Greek	827.3	920.3	780.4	**650.6**	749.8	687.9	634.4	**549.5**
Serbian	492.8	586.3	521.3	**408.0**	468.8	485.3	447.8	**365.7**
Russian	1656.8	2054.5	1920.4	**1466.2**	1609.5	1757.3	1648.3	**1309.1**
Italian	777.0	794.9	688.3	**592.2**	686.2	627.7	559.7	**493.4**
German	997.4	1026.0	1000.9	**831.8**	980.0	908.8	874.1	**761.5**
Finnish	1896.4	2438.8	2165.5	**1715.3**	1963.3	2233.2	2109.9	**1641.2**
Dutch	492.1	491.3	456.2	**381.4**	447.9	412.8	378.0	**330.1**
Japanese	1902.8	2662.4	2475.6	**1866.7**	1816.8	2462.8	2279.6	**1696.9**
Spanish	496.3	481.8	445.6	**387.7**	445.9	412.9	369.6	**331.2**

Table 4.1: Perplexity of language models trained on 8k and 16k sentences for different languages. *MKN* is an order 5 Modified Kneser-Ney language model. *LSTM* is a long short-term memory neural network language model with no pre-training. *CLWE* is an LSTM language model pre-trained with cross-lingual word embeddings, using English as the source language. *Interp.* is an interpolation of MKN with CLWE. Interpolated results gathered by Adam Makarucha at IBM using my models and code.

bilities with LSTM language model probabilities, which outperformed the separate language models.[6] Such interpolation allows for consistent improvement beyond the performance of MKN or CLWE-pre-trained LSTMs alone.

4.3.3 Other Target Languages

In Table 4.1 we present results of language model experiments run with other languages used as the low-resource source. In this table English is used in each case as the large target language with which to help train the CLWEs. The observation that the CLWE-pre-trained language model tended to perform best relative to alternatives

[6]The interpolated results were gathered by a collaborator, Adam Makarucha, using models and code I developed during my internship at IBM. They are included here as a point of comparison for the other results, which were entirely my own work.

at around 8k or 16k sentences in the English case prompted us to choose these slices of data when assessing other languages as sources.

The pre-trained LSTM language model outperforms its non-pre-trained counterpart for all languages. There is competition between *MKN* and the CLWE-pre-trained models. The languages for which *MKN* tends to do better are typically those further from English or those with rich morphology, making cross-lingual transfer of information more challenging. There seems to be a degree of asymmetry here: while all languages helped English language modelling similarly, English helps the other languages to varying degrees. This suggests the information is largely semantic, since the target languages vary in their divergence from English syntax (also, the window size used for training the CLWEs is large and will typically engulf the whole sentence). This, combined with the fact that English helps other languages to varying degrees, suggests that different source languages differ in the benefit they can reap from this technique. For all languages, interpolating MKN with the CLWE (*Interp.*) yields the best performance, corroborating the findings of Gandhe *et al.* (2014) and demonstrating that these methods are complementary.

Neural language modelling of sparse data can be improved by initializing parameters with cross-lingual word embeddings. The consistent performance improvements gained by an LSTM using CLWE-initialization is a promising sign for CLWE-initialization of neural networks for other tasks given limited source language data.

4.4 First Steps in an Under-Resourced Language

Having demonstrated the effectiveness of CLWE-pre-training of language models using simulation in a variety of well-resourced written languages, we proceed to a preliminary investigation of this method to a low-resource unwritten language, Na.

Yongning Na is a Sino-Tibetan language spoken by approximately 40k people in an area in Yunnan, China. It has no orthography and is tonal with a rich morphotonology. Given the small quantity of manually transcribed phonemic data available in the language, Na provides an ideal test bed for investigating the potential this method faces in a realistic setting, while highlighting its shortcomings. In this section we

æ˧ʂæ˧-tɑ˩mɣ̩˩ | ʈʂʰɯ˧ne˧-ʝi˥ pi˧-kɣ̩˩-tsɯ˩ | -mɣ̩˩!					
That's how the story is told!					
æ˧ʂæ˧-tɑ˩mɣ̩˩	ʈʂʰɯ˧ne˧-ʝi˥	pi˥	-kɣ̩˧˥	tsɯ˧˥	mɣ̩˩
story	like this	say	°abilitive	°rep	°affirm

Table 4.2: An example sentence from the Na corpus (sentence 137 from *The Sister's Wedding*), along with an English translation and glosses. Spaces delimit words, while hyphens delimit morphemes. The segmentation granularity that best facilitates dictionary lookup varies between words.

report results in Na language modelling and discuss hurdles to be overcome.

4.4.1 Experimental Setup

The phonemically transcribed corpus[7] consists of 3,039 phonemically transcribed sentences which are a subset of a larger spoken corpus. These sentences are segmented at the level of the word, morpheme and phonological process, and have been translated into French, with smaller amounts translated into Chinese and English. The corpus also includes word-level glosses in French and English. The lexicon of Michaud (2016) contains example sentences for entries, as well as translations into French, English and Chinese.

The lexicon consists of around 2k Na entries, with example sentences and translations into English, French and Chinese. Segmentation of the corpus is provided at the level of the morpheme, word and morphotonological process. However, the dictionary entries are mixed between morphemes and words without explicit distinction. Segmenting the corpus at only the word level or only the morpheme level yielded low hit rates, since the dictionary entries are distributed between morpheme-level entries and word-level entries. To choose an appropriate segmentation of the corpus, we used a hierarchical segmentation method where words were queried in the lexicon. If a given

[7]Available as part of the Pangloss collection at `lacito.vjf.cnrs.fr/pangloss`

	Types	Tokens
Tones	2,045	45,044
No tones	1,192	45,989

Table 4.3: The number of types and tokens across the Na corpus, given our segmentation method.

word was present then it was kept as a token, otherwise the word was split into its constituent morphemes and queried again. Without tones included, the total number of word types from the corpus found in the dictionary was 418/5980. The types not found were broken into constituent morphemes, where 400/1211 morpheme types in the corpus were found in the dictionary. Words and morphemes not found in the dictionary were given uniform random embeddings between -0.1 and 0.1. We evaluate perplexity at a mixed granularity, using units found with this dictionary-based segmentation approach. Table 4.2 shows an example sentence from the corpus, with glossing that motivates segmentation at both word level and morpheme level when appropriate.

We took 2,039 sentences to be used as training data, with the remaining 1k sentences split equally between validation and test sets. The phonemic transcriptions include tones, so we created two preprocessed versions of the corpus: with and without tones. Table 4.3 exhibits type and token counts for these two variations. In addition to the CLWE approach used in §4.2 and §4.3, we additionally tried lemmatizing the English Wikipedia corpus so that each token was more likely to be present in the Na–English lexicon. Lemmatization as a step makes sense for languages with varying syntax. For instance, it makes sense to conflate "run", "running" and "runs" to have the same vector representation on the English side since the equivalent Na word may not be inflected at all, or in a completely different way. By lemmatizing the English, we'd expect the embeddings to emphasize semantic relatedness over syntax. However, this may not be a good thing for language modelling, since language modelling is a syntax-oriented task.

	Tones	No tones
MKN	**59.4**	**38.0**
LSTM	74.8	46.0
CLWE	76.6	46.2
Lem	76.8	44.7
En-split	76.4	47.0

Table 4.4: Perplexities on the Na test set using English as the source language. *MKN* is an order 5 Modified Kneser-Ney language model. *LSTM* is a neural network language model without pretraining. *CLWE* is the same LM with pre-trained Na–English CLWEs. *Lem* is the same as CLWE except with English lemmatization. *En-split* extends this by preprocessing the dictionary such that entries with multiple English words are converted to multiple entries of one English word.

4.4.2 Results and Discussion

Table 4.4 shows the Na language modelling results.[8] Pre-trained CLWEs do not significantly outperform that of the non-pre-trained, and *MKN* outperforms both. Given the size of the training data, and the results of §4.3, it is no surprise that *MKN* outperforms the NNLM approaches. But the lack of benefit in CLWE-pre-training the NNLMs requires some reflection. We now proceed to discuss the challenges of this data to explore why the positive results of language model pre-training that were seen in §4.3 were not seen in this experiment.

Tones A key challenge arises because of Na's tonal system. Na has rich tonal morphology. Syntactic relationships between words influence the surface form tone a syllable takes. Thus, semantically identical words may take different surface tones than is present in the relevant lexical entry, resulting in mismatches with the lexicon.

If tones are retained, the percentage of Na tokens present in the lexicon is 62%.

[8]Note that although the perplexities are not phone-based, the tone of a word affects its orthographic representation. In the model without tones, tonal markers are discareded entirely and thus the calculated perplexities are different.

Removing tones yields a higher hit rate of 88% and allows tone mismatches between surface forms and lexical entries to be overcome. This benefit is gained in exchange for higher polysemy, with an average of 4.1 English translations per Na entry when tones are removed, as opposed to 1.9 when tones are present. Though this situation of polysemy is what the method of Duong *et al.* (2016b) is designed to address, it means the language model fails to model tones and doesn't significantly help CLWE-pre-training in any case. Future work should investigate morphotonological processing for Na, since there is regularity behind these tonal changes (Michaud 2008) which could mitigate these issues if addressed.

Polysemy It's known that many word embedding representations are limited in that they conflate different meanings of a word into a single vector (Camacho-Collados *et al.* 2016). We considered the polysemy of the tokens of other languages' corpora in the PanLex dictionaries. Interestingly, they were higher than the Na lexicon with tones removed, ranging from 2.7 for Greek–English to 19.5 for German–English. It seems the more important factor is the number of tokens in the English corpus that were present in the lexicon. For the Na–English lexicon, this was only 18% and 20% when lemmatized and unlemmatized, respectively. However it was 67% for the PanLex lexicon. Low lexicon hit rates of both the Na and English corpora must damage the CLWEs' modelling capacity.

Lexicon word forms Not all the forms of many English word groups are represented. For example, only the infinitive *'to_run'* is present, while *'running'*, *'ran'* and *'runs'* are not. The limited scope of this lexicon motivates lemmatization on the English side as a normalization step, which may be of some benefit (see Table 4.4). Furthermore, such lemmatization can be expected to reduce the syntactic information present in embeddings, which does not transfer between languages as effectively as semantics.

Some common words, such as *'reading'* are not present in the lexicon, while other words such as *'to_read_aloud'* are. Additionally, there are frequently entries such as *'way_over_there'* and *'masculine_given_name'* that are challenging to process.

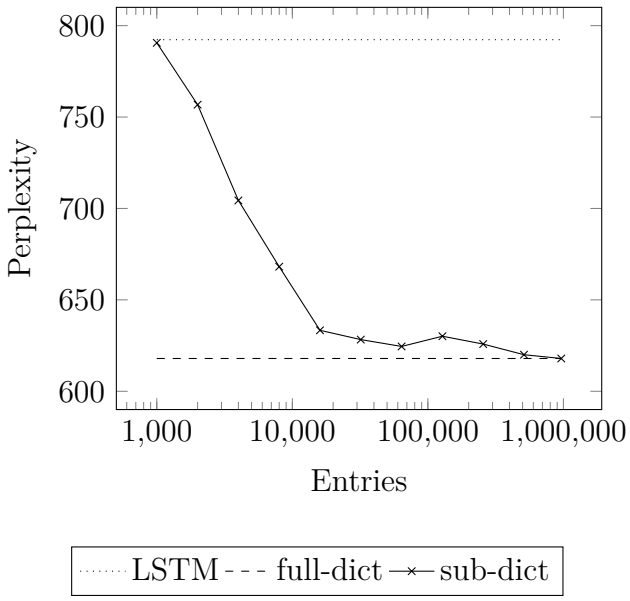


Figure 4.4: Perplexities of an English–German CLWE-pretrained language model trained on 2k English sentences as the dictionary available in CLWE training increases to its full size (*sub-dict*). As points of comparison, *LSTM* is a long short-term memory language model with no pre-training and *full-dict* is a CLWE-pretrained language model with the full dictionary available.

As an attempt to mitigate this issue, we segmented such English entries, creating multiple Na–English entries for each. However, results in Table 4.4 show that this failed to show improvements. More sophisticated processing of the lexicon is required, and is left to future work (§7.2.2).

Lexicon size There are 2,115 Na entries in the lexicon (a mix of morphs and lexemes) and 2,947 Na–English entries, which makes the lexicon especially small in comparison to the PanLex lexicon used in the previous experiments. Duong *et al.* (2016b) report large reductions in performance of CLWEs on some tasks when lexicon size is scaled down to 10k.

To better understand how limited lexicon size could be affecting language model performance, we performed an ablation experiment where random entries in the PanLex English–German lexicon were removed in order to restrict its size. Figure 4.4

shows the performance of English language modelling when training data is restricted to 2k sentences (to emulate the Na case) and the size of the lexicon afforded to the CLWE training is adjusted. This can only serve as a rough comparison, since PanLex is large and so a 1k entry subset may contain many obscure terms and few useful ones. Nevertheless, results suggest that a critical point occurs somewhere in the order of 10k entries. However, since improvements are demonstrated even with smaller dictionaries, this is further evidence that more sophisticated preprocessing of the Na lexicon is required.

Domain Another difference that may contribute to the results is that the domain of the text is significantly different. The Na corpus is a collection of spoken narratives transcribed, while the Wikipedia articles are encyclopaedic entries, which makes the genres very different. Since the tone and style are so different, it would be valuable in future work to use more closely matched corpora by using speech transcripts or narratives on the high-resource side.

4.5 Discussion

This chapter investigated a method for harnessing cross-lingual word embeddings for the purpose of improved language modelling in very low-resource contexts. Language modelling is a key component of systems for machine translation and speech recognition. Given methods for bilingual lexicon induction and word segmentation, we can relate the target and source languages and segment the words in some meaningful way. There is potential this method can then be used to learn a language model which can be used in a speech recognition system. Such language models may also facilitate rudimentary translation into the source language.

We argued in the discussion of the previous chapter (§3.4) that, by the time accurate manual transcriptions of any meaningful size have been acquired, linguists likely have already been working on lexicons and a small one may exist. It makes sense to harness this data as well. Importantly, it gives us the ability to harness vast quantities of monolingual information in English and other large languages in

order to improve processing of smaller languages, given that parallel text is always limiting. Thus, we are not actually dealing with bilingual audio, but rather bilingual information in the form of a lexicon.

This technique was not found to be effective for the threatened language scenario of Na. However, in light of the results in the other languages examined, the technique could find more applicability in a low-resource language that has a more comprehensive dictionary and a limited web presence for scraping (see Gauthier *et al.* (2016)).

4.5.1 Future Work on Na Language Modelling

The technique doesn't work out of the box for Na, setting a difficult and compelling challenge of harnessing the available Na data more effectively. The lexicon is a rich source of other information, including part-of-speech tags, example sentences and multilingual translations. In addition to better preprocessing of the lexical information we have already used, harnessing this additional information is an important next step to improving Na language modelling. The corpus includes translations into French, Chinese and English, as well as glosses. Some CLWE methods can additionally utilize such parallel data (Coulmance *et al.* 2015; Ammar *et al.* 2016) and we leave to future work incorporation of this information as well. The tonal system is well described (Michaud 2008), and so further Na-specific work should allow differences between surface form tones and tones in the lexicon to be bridged.

Though the results for language modelling in Na are inconclusive, the ablation experiment, where the dictionary size was reduced significantly in order to assess how it affected the model's performance, suggests the method can still be useful with limited dictionary sizes. Importantly, very low dictionary sizes never adversely affected language modelling performance, which suggests that this method is worth trying and unlikely to hurt performance. In the specific context of Na, there were a variety of confounding factors which probably played a role in preventing distributional information from the English corpus from helping Na language modelling.

4.5.2 Beyond Na

Our results corroborate the observation that MKN performs well on rare words (Chen *et al.* 2015). Interpolation is an effective means to harness this strength when training data is sparse. Furthermore, hybrid count-based and NNLMs (Neubig and Dyer 2016) promise the best of both worlds for language modelling for low-resource languages.

In order for this approach to be applicable in low-resource language modelling for tasks such as speech recognition and machine translation, where there will frequently be out-of-vocabulary words, a promising line of work is to integrate character and phoneme level information into the language model (Lankinen *et al.* 2016; Verwimp *et al.* 2017) to help the model cope with sparsity at the word token level.

So far in the thesis we have considered two language documentation contexts. The first is where we have small quantities of unsegmented phonemic transcripts in parallel with translations in a larger language. Such data can arise when a linguist carefully transcribes monolingually without the need for deep knowledge of the language. We demonstrated in Chapter 3 that bilingual lexical items can be learnt effectively but argued against measuring the quality of a bilingual lexicon intrinsically. The context considered in this chapter is one where we again have small quantities of accurate source transcriptions, however we additionally assume a bilingual lexicon, which may have been gathered by a linguist in the process of transcription, or by automatic methods in a similar vein to those of Chapter 3. In this context however, we demonstrate that there is no requirement of parallel data in order for improved language modelling to be achieved. This absence of this requirement allows for the true magnitude of data in large languages to be harnessed in improving modelling of smaller ones. In this work we used 5 million sentences of Wikipedia data in the large language. If this method is pursued it would be wise to harness even more data, and to harness from multilingual (more than two languages) contexts as per the work of Duong *et al.* (2017).

We now proceed to a third context, where we removed the assumption of accurate phonetic transcriptions, while maintaining the constraint of limited data. By remov-

ing the two key assumptions of ample data and of correct data, we now model when restricted by the constraints that are very real and limiting in language documentation. In doing so, we make methods available to a much larger spectrum of languages at different levels of documentation.

Chapter 5

Harnessing Translations for Improved Phoneme Transcription

Large portions of this chapter have appeared in the following papers:

> Adams et al. (2016) Learning a translation model from word lattices, in *17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, California, USA.
>
> Adams et al. (2016) Learning a lexicon and translation model from phoneme lattices, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA. pp. 2377–2382.

5.1 Introduction

The previous chapters have assumed the availability of correct phonemic transcriptions of source language speech, which is only possible when speech has been manually transcribed by a trained linguist. However, such manual phonemic transcription is costly and can become a bottleneck in linguistic fieldwork. To illustrate, one minute of speech takes around one hour for an expert to phonetically transcribe (Do *et al.* 2014a). In order for a radical speedup in the language documentation process to be possible, effective automatic phoneme recognition must be achieved, thus improving

the efficiency of linguists documenting the language. This chapter and Chapter 6 address this problem.

Without the aid of a lexicon or language model trained on abundant text, automatic phoneme transcription is an error-prone affair. Because of this, important benefits lie in harnessing cross-lingual information via multilingual acoustic models or via translations of speech. This chapter explores methods that harness translations of the speech in order to improve automatic phoneme transcription quality. Can translations of speech improve transcription of that speech, even when a prior translation model relating the languages is not known? We explore methods to learn such a translation model to aid in speech recognition, in the process additionally inferring word segmentation and lexicons. Insights from Chapter 3 motivate evaluation of bilingual lexicons through their ability to help in the extrinsic task of phoneme transcription. Thus quantitative results in this chapter are measured with phoneme error rate (PER) and word error rate (WER) of transcription output. Beyond this quantitative analysis, we also present qualitative analysis of lexicons and translation models underlying the model.

The underlying idea common to the methods explored in this chapter is that speech recognition should become an easier task when a translation is available to help disambiguate what words might have been spoken. Figure 5.1 illustrates this using a toy German–English example. This concept of harnessing translations to improve speech recognition has a long history of application in computer-aided translation (see 2.4.2). There are two key distinguishing features of the work in this chapter from previous work. Firstly, previous work uses translation models trained on text as prior information, but in our case the translation models are trained on inaccurate representations of the speech we want to transcribe: either error-prone 1-best transcriptions or lattices. Secondly, the translation model is trained with word-segmented text, whereas in this chapter we explore training it on unsegmented phoneme sequences, as well as phoneme lattices from real automatic speech recognition (ASR) systems.

§5.2 begins with a preliminary investigation into the use of phoneme classes to facilitate translation modelling between source phonemes and target translations even when the exact form of the phonemes is unknown. This exploration highlights short-

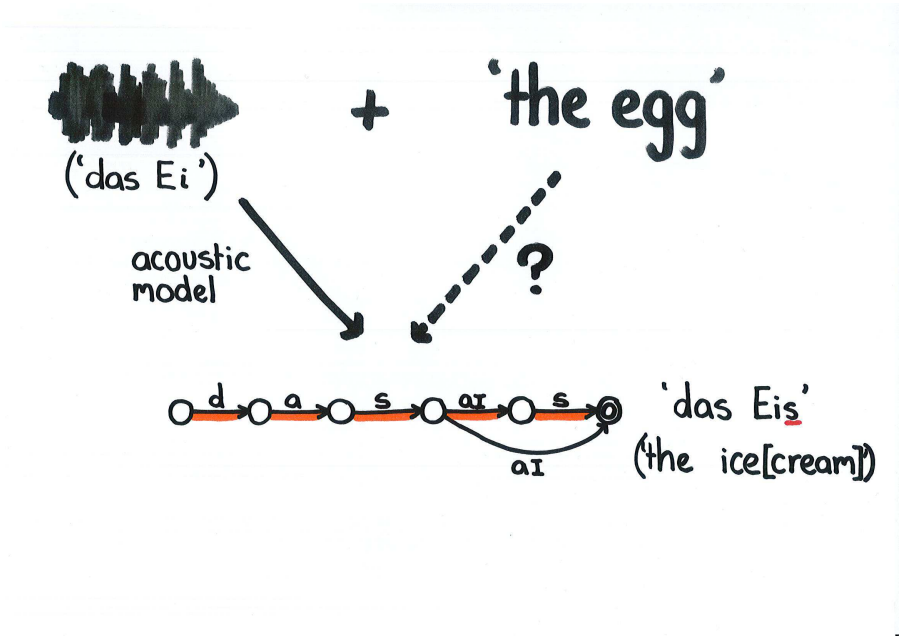


Figure 5.1: Top: a spoken utterance in German, along with a translation that should help in finding the correct transcription. Bottom: a phoneme lattice output of an acoustic model. Using an acoustic model we can construct a lattice containing phoneme sequences that likely explain the acoustic signal, but frequently the most probable path (in orange) will be incorrect. The available English translation should guide the transcription away from 'das Eis' and towards 'das Ei.' Without prior translation model probabilities, translation relationships must be learnt in the face of such error-prone representations of speech.

comings of the model, motivating development of a subsequent model to address these issues, described in §5.3 and §5.4. This second model takes as input phoneme lattices and sentence-level orthographic translations, jointly segmenting paths in the lattices and aligning the implied words with an orthographic translation while learning a translation model. This allows it to find a better transcription in the original lattice even when no prior information relates the languages. As a stepping stone towards developing the second model, we produce a word-based model that aligns without segmenting in order to first test the hypothesis that lattice alignment can harness

bilingual information for better speech recognition.

5.2 Phrase Alignment Using Phoneme Classes

This section (§5.2) describes preliminary work in the alignment of phoneme *classes*. Shortcomings of the model leave it unable to reduce transcription errors. Readers interested in more effective models should feel comfortable skipping to §5.3.

This exploration is motivated by the observation that humans have a tendency to misidentify phonemes that fall within a common class. Consider that an /s/ in *sight* is more likely to be heard as another fricative (eg. the /f/ in *fight*), than as sonorant (eg. /m/ as in *might*). We explore the use of categorizing phonemes into classes where the member phonemes are likely to be confusable with one another. For this we first used a linguistically motivated taxonomy, since it groups phones by articulatory and acoustic properties which tend to lend themselves to confusion.

Table 5.1 shows three partitions of the phoneme set based on a linguistically motivated taxonomy. We chose three different granularities for fragmenting German phonemes. Note that the transition from the fine partition to the coarse partition is not agglomerative—the phonemes in the plosives class in the `Fine` partition are then spread across the voiced and voiceless consonants class in the coarse partition.

Generalizing Phonemes to Classes for Machine Translation

To motivate the use of such classes, consider the performance of phoneme–word machine translation systems[1] when phonemes are replaced with symbols representing the coarser phoneme classes of Table 5.1. Such a configuration models a scenario where accurately determining the specific phoneme is error-prone, but determining a more general class is reliable. Table 5.2 illustrates the replacement of an international phonetic alphabet (IPA) representation of *Maus* with class tokens of Table 5.1, which is then used as source-side training data in the machine translation system.

[1]The models were trained and evaluated with a similar architecture and experimental setup as the one described in §3.2, except with ~457k training sentences.

	`Fine`
Plosives	p b t g d k ʔ
Affricates	pf ts tʃ
Fricatives	f v s z ʃ ʒ ç j x h
Sonorants	m n ŋ l ʁ r
Checked vowels	u ɪ ɛ a ɔ ʊ ʏ œ o i
Schwa-like	ə ɐ
Free vowels	iː eː ɛː aː oː uː yː øː
Free diphthongs	aɪ aʊ ɔʏ
	`Medium`
Plosives	p b t g d k ʔ
Fricatives	f v s z ʃ ʒ ç j x h pf ts tʃ
Sonorants	m n ŋ l ʁ r
Vowels	u ɪ ɛ a ɔ ʊ ʏ œ o i ə ɐ iː eː ɛː aː oː uː yː øː aɪ aʊ ɔʏ
	`Coarse`
Voiced consonants	b g d v z ʒ j m n ŋ l ʁ r
Voiceless consonants	p t k ʔ pf ts tʃ f s ʃ ç x h
Vowels	u ɪ ʊ ʏ i ɛ a ɔ œ o ə ɐ iː u yː eː ɛː aː oː øː aɪ aʊ ɔʏ

Table 5.1: Three taxonomical groupings of phoneme classes of different granularities, from fine to coarse.

Written	Maus
IPA	m aʊ s
`Fine`	sonorant free-diphthong fricative
`Medium`	sonorant vowel fricative
`Coarse`	voiced-consonant vowel voiceless-consonant

Table 5.2: Representing phoneme sequences with phoneme classes.

Source tokens	BLEU
Phonemes	20.77
`Fine`	20.73
`Medium`	16.11
Coarse	12.48

Table 5.3: German–English machine translation results when phonemes are replaced with symbols denoting classes.

Table 5.3 shows end-to-end machine translation results where instead of translating German phonemes to English words, the German phonemes were replaced with tokens representing broader phonemic classes. Interestingly, using the `Fine` class groups yields no substantial change in the machine translation performance, though there is large degradation when coarser classes are used. This highlights the information redundancy in natural language and suggests that phrase alignment can be resilient to phonemic ambiguity. Since alignment is still feasible when some amount of phonemic information is removed, perhaps English translation contexts can help guide restoration of specific phonemic information.

5.2.1 Resolving Transcription Errors: Method

We now proceed to the task of harnessing translations to resolve transcription errors. That is, given a phoneme transcription replete with errors and a corresponding word-level translation, can this translation guide us to a more correct phonemic transcription by somehow relating subsequences of the phonemic transcription with the translation? The first, preliminary approach to this task that we discuss is motivated by this observation that replacing phonemes with symbols to denote classes can retain much of the information in the phoneme sequence useful for the purpose of translation. The advantage this approach might have over finding the one-best path in a phoneme lattice is that information in translations can inform how the

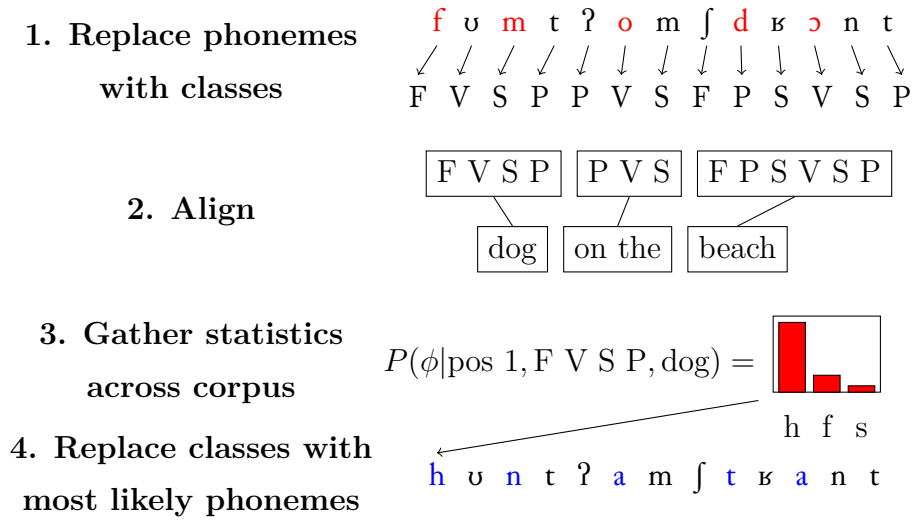


Figure 5.2: An illustration of the error resolution process for a phonemic transcription of *Hund am Strand*. Erroneously transcribed phonemes are indicated in red. Phoneme classes are represented in upper-case, where F represents fricatives, V represents vowels, S represents sonorants, and P represents plosives.

transcription is realized. The method has four steps:

1. Replace transcribed (frequently erroneous) phonemes with tokens representing broader phonemic classes.

2. Align the phoneme class tokens to the English translations.

3. Estimate probabilities of phonemes occurring given classes and translation contexts.

4. Replace phoneme class tokens with the maximum likelihood phoneme given the alignment context.

Figure 5.2 illustrates visually using a toy example the process of resolving errors in a phonemic representation of the German phrase "Hund am Strand." More formally, given a corpus of phonemic transcriptions of utterances X (possibly the output of a speech recognition system) and corresponding English translations E, we aim to find a set of more accurate transcriptions $\hat{X}$ without supervision.

Step 1: Generalize phonemes to classes Consider a partition of phoneme sets $P = \langle p_i \dots p_I \rangle$ representing the chosen grouping of phoneme classes. For each phonemic transcription $\boldsymbol{x} \in X$, we convert each phoneme $x \in \boldsymbol{x}$ into a label representing p_i, where p_i is the set in the partition P that contains x, yielding a phoneme class representation of the sentence, which we denote as $\boldsymbol{c}$. Do this for each sentence, yielding an ordered sequence C where each $\boldsymbol{c} \in C$ corresponds to a sentence in the same fashion as X and E.

Step 2: Phrase alignment Phrase align C and E to arrive at alignments A such that each $\boldsymbol{a} \in A$ describes a collection of many-to-many alignments for a sentence pair, where each $a \in \boldsymbol{a}$, aligns a subsequence of phoneme classes $c_i \dots c_j$ to sequences of words in the English sentence $e_k \dots e_l$.

Step 3: Estimating phoneme probabilities given alignment contexts By considering each $x \in \boldsymbol{x} \in X$, their corresponding class c, and alignment context a, we can estimate a categorical distribution over phonemes given that context, $P(x|c, a)$. For example, in Figure 5.2 we can estimate the probability of different fricatives taking the place of the first phoneme class 'F' in a source phrase pattern 'F V S P' that is aligned to 'dog.'

Step 4: Take the most likely phoneme given each context To produce the output phoneme stream, the maximum likelihood phoneme is taken for each phoneme class label and associated alignment context, $\text{argmax}_x P(x|c, a)$.

5.2.2 Experimental Setup

In order to imitate the situation where we have inaccurate automatic speech recognition output aligned at the phrase level to English text, yet in significant quantities, we introduced artificial errors into a dataset similar to the one used in the phoneme-based machine translation experiment described in §3.2.2.

Data

The data used is the German–English corpus from shared task for the Association for Computational Linguistics (ACL) 2005 Workshop on Building and Using Parallel Texts. This corpus consists of 457,079 sentences, a subset of Europarl (Koehn 2005). In order to simulate automatic phoneme transcription data of a language with no orthography, we first convert the German data into a sequence of phonemes using the MARY text-to-speech system (Schröder and Trouvain 2003). This training set is thus a superset of the training data used in §3.2.

Error Simulation

We add artificial errors into the phoneme corpus. This process was parameterized by the parameter ρ, the probability of any given phoneme being erroneously substituted for another phoneme within its class. For each phoneme in the phoneme corpus, the phoneme was substituted by another phoneme in its class with probability ρ and remained the same otherwise. When substituted, it uniformly took the form of each other phoneme in its class c, with probability $\frac{1}{|c|}$. Thus ρ is also a close approximate measure of the phoneme error rate of the phoneme corpus. For error generation, the `Fine` partition was used, but for error resolution all three partitions were evaluated. 5 different values of ρ were used, between 0.1 and 0.5, representing varying transcription accuracies.

Phrase Alignment

We used the Bayesian inversion transduction grammar model as implemented in Pialign (Neubig *et al.* 2011b; Neubig *et al.* 2012a), since it was the best performing approach in Chapter 3, able to effectively deal with phoneme granularities that cannot be done using conventional word-level alignment and heuristic phrase extraction.

Phoneme error rate (%)			
ρ	`Fine`	`Medium`	`Coarse`
10.0	**7.3**	12.2	12.4
20.0	**13.2**	**18.0**	20.5
30.0	**19.6**	**24.4**	**29.2**
40.0	**27.0**	**31.5**	**38.4**
50.0	**36.2**	**40.3**	**48.2**

Table 5.4: Phoneme error rates of the output of the error resolution algorithm using different granularities of phoneme classes. ρ is the phoneme error rate of the inaccurate corpus, before resolution. `Fine`, `Medium` and `Coarse` refer to the granularity of the phoneme classes used by the resolution algorithm (see Table 5.1). Bold scores indicate scores that reduced the error rate.

5.2.3 Results

Table 5.4 shows the phoneme error rate across the corpus when this method is applied. In most of the experiment configurations, this method yielded an output phoneme sequence with a lower phoneme error rate than the inaccurate baseline. The method makes the most meaningful improvements in the phoneme error rate when the phoneme classes used are `Fine`, followed by `Medium` and then `Coarse`, with three configurations underperforming the baseline between the `Medium` and `Coarse` partitions.

When improvements do occur using the `Coarse` partition, they are not significant. This is understandable, since the scope of the possible confusion is greater with coarser classes, and there is less information that Pialign has to operate with in the alignment phase. Additionally, since the artificial errors were introduced using the `Fine` partition, these results suggest that when the classes used to resolve the errors are very different to those used to generate the errors, this approach to error resolution can become ineffective (though it is more robust to smaller changes, as

evidenced by the results using the `Medium` partition).

5.2.4 Discussion

These results suggest translations can be harnessed to improve phoneme transcriptions in an unsupervised fashion in a constrained scenario where errors conform to clear discrete phoneme classes. Though these results show that alignment of source coarser phoneme classes to English words can be effective, and that the general concept of harnessing translation modelling in uncertain contexts is promising, several issues make more general applicability of this specific approach questionable.

The artificial errors generated assume that errors manifest only as substitutions according to the presupposed phoneme classes. For this technique to be applicable to improving speech recognition, experiments should be run on data with realistic speech recognition errors. In §5.2.5 below, we more accurately model errors in the input training data.

A related issue in error generation is that spurious insertions and deletions of phonemes in the transcription are not modelled. However, they are prevalent in real speech recognition errors (see Table 5.7). This suggests that there is a fundamental issue with the notion of phoneme class: real transcription errors cannot be modelled with substitutions within phonemic classes.

In the remainder of this chapter we address both of these issues. Before moving to a model capable of addressing insertions and deletions on real speech (§5.3 and §5.4), we first assess how the approach we have described so far can handle more realistic substitution errors that fall outside of the phoneme classes.

5.2.5 Improved Error Simulation and Phoneme Confusion Modelling

The key issue of the above method is that error generation and resolution use unrealistic phoneme categorization, so in this section we make two changes. Firstly, we instead model errors from real speech recognition data and use that to generate ar-

tificial phoneme substitution errors based on actual confusion rates. Secondly, rather than using taxonomically motivated classes, we use an agglomerative clustering approach to create classes based on the frequency of confusion between phonemes in real data.

Error Simulation

In order to more accurately model error generation, we procured 3,986 sentences of German automatic speech recognition output from a Quaero HMM-GMM system (Stüker *et al.* 2012). This data was then used to estimate phoneme substitution rates via minimum edit-distance alignments with gold transcriptions, before random sampling was used to add artificial noise into the same training data from before, with similar error rates and proportions as in the real data. As a first step, we only explore substitution errors. The system had a phoneme error rate of 43.9%.

Phoneme Clustering

Previously we had considered a linguistic taxonomy of phonemes in order to determine these classes, but actual phoneme substitutions in the ASR data indicated that these taxonomies do not necessarily yield the best phoneme classes and so we empirically ground the categorization.

We performed agglomerative hierarchical clustering such that phonemes more confusable with one another were merged into common classes. The distance between two phonemes p and q is the reciprocal of the number of times p had been substituted for q or q had been substituted for p. We applied agglomerative hierarchical clustering, using average linkage criteria. Average linkage was chosen since for our purposes it is desirable to have classes where each phoneme is confusable with one another, minimizing cases where two phonemes in a given class are rarely confusable.

We chose the configurations that gave 2 classes, 5 classes and 12 classes per group as representative of varying granularities. These are shown in Table 5.5. The reasoning for these class group choices was as follows. `Coarse` (2 classes) is the group that is able to account for as much possible phoneme confusability as possible while

	`Fine`
Class 0	ɔʏ
Class 1	øː
Class 2	aʊ
Class 3	j iː
Class 4	s z ts
Class 5	ŋ m n
Class 6	ʊ uː ɔ oː
Class 7	b d g k p t v
Class 8	a aː œ aɪ
Class 9	ɛ yː ɪ eː ʏ ɛː ə
Class 10	h ʁ l ɐ f
Class 11	x ç ʃ
	`Medium`
Class 0	ɔʏ
Class 1	aʊ
Class 2	ç, x
Class 3	b ʃ g f n h k ɐ m ts ŋ p s t v ʁ z l d
Class 4	a ɛ eː oː aɪ aː œ j ʏ ɔ yː ɪ øː ʊ uː ɛː iː ə
	`Coarse`
Class 0	b ʃ g f n h k ɐ m ts ŋ p s ç t v x ʁ z l d
Class 1	a ɛ eː oː aɪ aː œ j ʏ ɔ yː ɪ øː ʊ uː ɛː ɔʏ aʊ iː ə

Table 5.5: Groupings of phonemes at different granularities based on hierarchical agglomerative clustering.

Method	PER
Baseline	32.57
`Fine`	36.46
`Medium`	46.95
`Coarse`	44.79

Table 5.6: Phoneme error rates when applying the error resolution method with different granularities for phoneme classes. Phoneme error rates all increase over the baseline (of no error resolution), despite the inclusion only of substitution errors in the synthetic data.

still yielding information. Groupings more granular than `Fine` (12 classes) rapidly degenerated towards a situation where each phoneme is of its own class.

5.2.6 Results and Discussion

Table 5.6 shows the results of error resolution when simulated substitution errors mimic that of real data and agglomerative clustering is used to determine phoneme classes. Regardless of the granularity of the class used, the approach underperforms the baseline. Unsurprisingly, `Medium` performs the worst, since it is distinguished from `Coarse` only by the creation of two single-phoneme classes, and one class with two phonemes, which are largely unable to account for phoneme misclassifications.

More realistic simulation of substitution errors demonstrates that categorization of phonemes is an ineffective model, even when categories are motivated empirically by observations of a real speech recognition system. Phoneme substitution errors crossing class boundaries negatively impacts alignments, since the source representation of a word (as class tokens) varies. This has the effect that, in addition to those specific substitution errors being unable to be resolved, the alignments propagate the damage to surrounding phonemes grouped together in the alignment.

There are various issues in terms of both the experimental setup and the modelling

Error type	% of total
Substitutions	66.26
Deletions	21.66
Insertions	12.08

Table 5.7: The proportions of different types of errors in the German ASR data.

approach. Firstly, simulated data is still used, though it is better than that of the previous section as it exhibits similar substitution error rates and types as those from a speech recognition system. This data is problematic in that a) it is phonemic data of European parliamentary proceedings, thus the modality is inconsistent with real speech and effects of features such as coarticulation are limited by the quality of the text-to-speech system producing the phonemic representation; b) a generous amount of data is used, which is unlikely to be attainable in a language documentation setting (~450k sentences); c) insertions and deletions are not modelled. Insertions and deletions can be addressed in a similar unigram model as the substitutions, but it would be better yet to use real speech recognition output as we do in §5.3 and onwards.

As for the model, this experimental evidence suggests that phonemic categories are limited in their capacity to model phoneme confusion owing to deviation from these categories. Because the experimentation doesn't use real speech recognition output, the model also fails to harness further information an acoustic model could yield, such as lattices which describe alternative phonemic transcriptions to explain the observed sounds.

In retrospect, the data in Table 5.8 strongly suggests it is ineffective to use a method relying on phoneme substitution errors consistent with disjoint classes. In the best case, 32% of phonemes will cross class boundaries (in the case of coarse-grained clustering, assuming only substitution errors). Assuming uniformity of this statistic across phonemes, there's a $(1 - 0.32)^4 \approx 0.22$ probability a given 4-phoneme word will be converted to a class-token representation consistent with the underlying true phonemes, as required for correction. For longer words, this probability quickly drops

Granularity	Taxonomical	Clustering	Clustering (subs-only)
`Coarse`	35.0	45.9	68.1
`Medium`	31.4	42.9	64.7
`Fine`	16.2	23.5	35.5

Table 5.8: The percentage of phoneme transcription errors in the speech recognition data that were consistent with taxonomical or clustering-based phonemic classes (of three different granularities: `Coarse`, `Medium` and `Fine`). The right-most column restricts sources of errors to substitutions.

towards 0. This back-of-the-envelope calculation and consideration of its implications should have been done when clustering was first performed the first time, but it was not.

Another limitation is that experimental results have the limitation that spoken speech takes a different form to written speech converted to phonemes, with phenomena such as coarticulation present. In using a phonemic representation of text, our model of true speech is thus limited in capacity to that of the speech synthesizer.

5.3 Learning a Translation Model from Word Lattices

In this section and the next (§5.4) we address the key problems of the work presented earlier in this Chapter. Rather than error simulation, we use real output from a speech recognition system. We stop using the notion of phoneme confusion classes, which does not accurately reflect the real nature of recognition errors. Rather than a 1-best transcription, we use a lattice which conveys more information. Rather than the assumption of large amounts of data (~450k sentences), we evaluate on small amounts of data realistic in the language documentation context (as little as 1 hour of bilingual speech). Despite these changes, the underlying idea remains: using

information available in translations to help automatic transcription.

In the rest of this chapter we use lattice representations of speech. Lattices contain more information than a 1-best transcription. This information, along with translations, can be harnessed to find more accurate transcriptions (ie. paths through the lattice). Lattices can represent varying length sequences, implicitly allowing modelling of insertions and deletions, overcoming a key shortcoming of the model in §5.2. Composition of lattices with finite-state transducers can be used to express and infer translation models, thus modelling the acoustic signal and the translation in a cohesive probabilistic framework. In this section we first explore the effectiveness of this concept at the word level, by assuming we have a lexicon available with which to make word lattices. Then in §5.4, we generalize the approach to learn a lexicon from phoneme lattice input for the case when a comprehensive lexicon is not available in the language being documented.

Translation models have been used in prior work to improve automatic speech recognition when speech input is paired with a written translation, primarily for the task of computer-aided translation, where a human translator speaks their translation of a written document (Brown *et al.* 1994; Vidal *et al.* 2006; Khadivi and Ney 2008; Reddy and Rose 2010; Pelemans *et al.* 2015). Existing approaches require large amounts of parallel text for training the translation models, a scarce resource for most language pairs even when each language has substantial monolingual data. We propose a model for learning lexical translation parameters directly from the word lattices for which a transcription is sought. The model is expressed through composition of each lattice with a weighted finite-state transducer representing the translation model, where inference is performed by sampling paths through the composed finite-state transducer. We show consistent word error rate reductions in two datasets, using between just 20 minutes and 4 hours of speech input, always outperforming a translation model trained on the 1-best path.

Beyond the connection to computer-aided translation, the work in this section and section 5.4 has parallels with topics described in Chapter 2, including speech translation, where speech lattices are composed with translation models (Casacuberta *et al.* 2004; Matusov *et al.* 2005), translation modelling from automatically transcribed

speech, which has previously used 1-best transcriptions (Paulik and Waibel 2013), and Bayesian word alignment (Mermer *et al.* 2013; Li *et al.* 2013).

We use a generative model that assumes the acoustic signal and written translation are produced by some underlying word sequence we seek to recover. This model is expressed by composing a word lattice that expresses information from the acoustic and language models with a weighted finite-state transducer (WFST) that expresses lexical translation probabilities constrained by the observed translation (see Figure 5.3). These parameters are learnt by sampling paths through the composed WFST, which corresponds to sampling a word sequence and its alignment to the written translation. A likely source sentence is recovered by finding the shortest path in the WFST.

In experiments on the Fisher and CALLHOME Spanish–English Speech Translation Corpus (Post *et al.* 2013), we compare word error rates with those of ASR 1-best paths and a stronger baseline that trains an existing translation model on 1-best recognition results. The distinction between these two methods is that the former uses only monolingual information in a 1-best transcription, whereas the latter harnesses the 1-best transcription and translations in order to learn a translation model and subsequently improve the 1-best transcription. We demonstrate reduced word error rates of 4.1% to 5.6% relative over the 1-best paths, and also show 2.3% to 2.4% relative improvement over the alternative model that uses parameters learnt from 1-best transcriptions. These results indicate that the mere existence of translations of what is to be transcribed can help with ASR. Moreover, it shows promise for models of this type for computer-aided translation and also for speech recognition for low-resource languages, where neither translation nor recogniser technologies are currently adequate.

The method differs from previous work in that (a) it depends on no parallel text training data, and that (b) the translation model is trained directly from word lattices to harness more information than is available in the 1-best ASR hypothesis alone.

This approach uses techniques similar to those found in the Bayesian word alignment literature (Mermer and Saraçlar 2011; Mermer *et al.* 2013; Li *et al.* 2013). However, rather than sampling alignments between observed source and target word

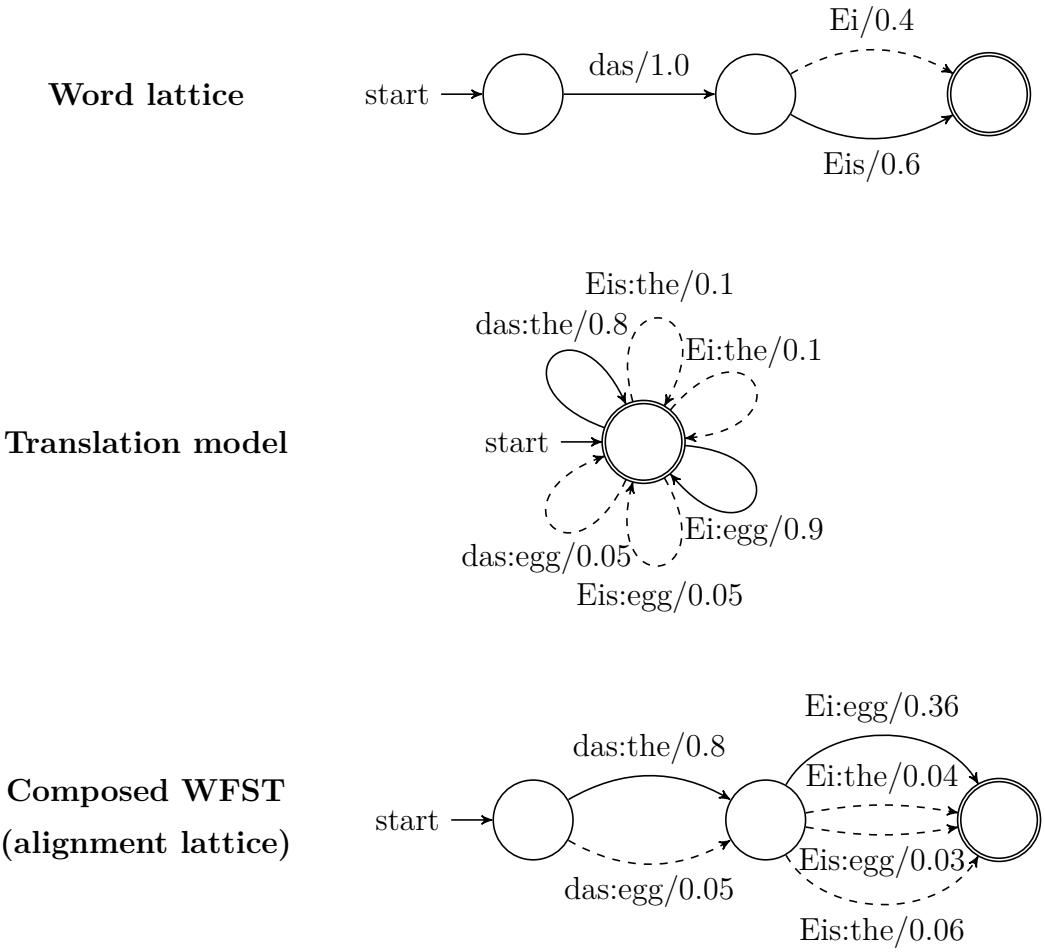


Figure 5.3: Illustration of the components of the WFST architecture. The alignment lattice is the result of composing the word lattice with the translation model which is constrained based on an observed English transcription "the egg." In this way, translation model probabilities can guide transcription. In our formulation, only the word lattice probabilities are given; the translation model parameters are learnt.

sequences, we sample paths through the source word lattice jointly with alignments to the target (translation) word sequences. This approach is also similar to that of Neubig *et al.* (2012a), where a lexicon and language model are learnt directly from phoneme lattices. However, rather than composing a phoneme lattice with a lexicon and a language model WFST, we compose a word lattice with a WFST representing a translation model.

5.3.1 Model Description

ASR Lattices

ASR is characterized by the search problem

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmax}}\, P(\boldsymbol{x}|\boldsymbol{f})P(\boldsymbol{f}) \tag{5.1}$$

where $\boldsymbol{f}$ represents an unobserved sequence of words $f_1 \dots f_J$ that produced the sequence of observed acoustic features $\boldsymbol{x} = x_1 \dots x_T$, and $\hat{\boldsymbol{f}}$ is our best guess of those words. Note that the length J is unknown ahead of time.

An ASR lattice encodes multiple transcription hypotheses, as shown in Figures 5.3 and 5.4, where each edge corresponds to a word f_i. The acoustic model (AM) and language model (LM) probabilities, $P(\boldsymbol{x}|\boldsymbol{f})$ and $P(\boldsymbol{f})$ respectively, are captured by the weights of the edges.

For any path $\boldsymbol{f}$ through the lattice, its probability can simply be determined with $P(\boldsymbol{f}) = \prod_{j=1}^{J} P_L(f_i)$, where $P_L(f_i)$ is the probability of the ith edge in that path, where the language model probability is assumed to factorize with the graph.

The most likely path $\hat{\boldsymbol{f}} = \operatorname{argmax}_{\boldsymbol{f}} P(\boldsymbol{f})$ can be determined by finding the shortest path through the lattice in question using Dijkstra's algorithm, if probabilities are represented as negative log probabilities.

Proposed Model

The proposed model also uses translation models to aid in ASR by incorporating additional information in the form of an observed sequence $\boldsymbol{e} = e_1, \dots, e_I$ of translated

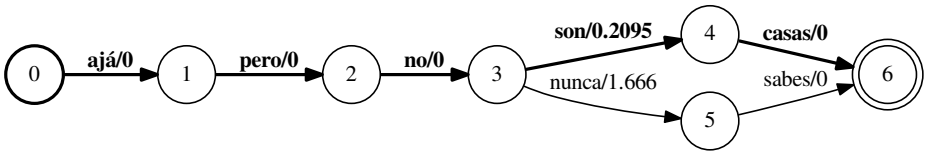


Figure 5.4: A lattice obtained from speech recognition of *'Ahá, pero una nunca sabe'* (crowdsourced translation: *'Aha, but one never knows'*), with negative log probabilities. Note that the gold transcription cannot be found in the lattice. The probability of *'son'* is incorrectly higher than the probability of '*nunca*.'

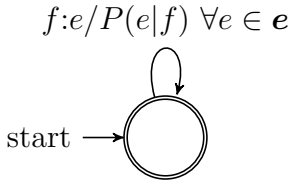


Figure 5.5: Reduced translation model template. Edges are added to the WFST only if the word e is present in the utterance-level written translation $\boldsymbol{e}$.

words and alignments $\boldsymbol{a}$ between $\boldsymbol{f}$ and $\boldsymbol{e}$. If $P(\boldsymbol{e}, \boldsymbol{a}|\boldsymbol{f})$ is factored into the search problem, then we can reframe the problem as

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\text{argmax}} \sum_{\boldsymbol{a}} P(\boldsymbol{x}|\boldsymbol{f}) P(\boldsymbol{f}) P(\boldsymbol{e}, \boldsymbol{a}|\boldsymbol{f}), \tag{5.2}$$

which equals $P(\boldsymbol{e}, \boldsymbol{x}, \boldsymbol{f})$ under the reasonable assumption that $\boldsymbol{x}$ and $\boldsymbol{e}$ are conditionally independent given $\boldsymbol{f}$ (ie. $\boldsymbol{f}$ contains all information pertinent to translation).

This problem can be reduced to a similar shortest-path problem as with traditional ASR. This is done by composing our original lattices with a WFST that represents translation model probabilities, as shown in Figure 5.5. The resulting composition of the final lattice and the constrained translation model can be seen in Figures 5.3 and 5.6.[2]

[2]For the examples in the figures and this formulation, we disregard the possibility of the null

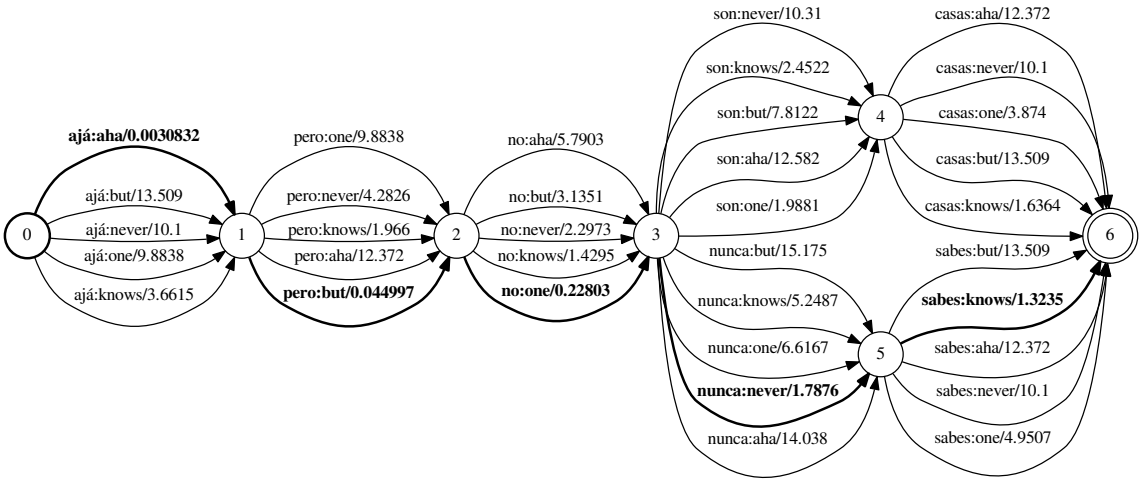


Figure 5.6: The lattice composed with the translation model WFST. Each edge for a given Spanish word is replaced with a set of edges that transduce to different English words with probabilities re-weighted by the translation model. Note that *'nunca'* is now correctly given more weight than *'son'*, given the added information of the English translation 'never.'

In this framework, each path represents a sequence of source tokens $\boldsymbol{f}$ and their alignments $\boldsymbol{a} = a_1 \dots a_J$ between tokens in $\boldsymbol{f}$ and types in $\boldsymbol{e}$. However, this generative story can fail to describe the observed reference translation $\boldsymbol{e}$ if, say, all words in $\boldsymbol{f}$ align to the first word in the translation, e_1. This is a *deficient* model of $P(\boldsymbol{e}, \boldsymbol{a}|\boldsymbol{f})$. An alternative formulation is to condition on $\boldsymbol{e}$, ie. using $P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e})$. Since we assume language model probabilities $P(\boldsymbol{f})$ are already included, this gives rise to a product model where probabilities over $\boldsymbol{f}$ are given by two different sources:

$$\hat{\boldsymbol{f}} = \operatorname*{argmax}_{\boldsymbol{f}} \sum_{\boldsymbol{a}} P(\boldsymbol{x}|\boldsymbol{f}) \frac{P(\boldsymbol{f})P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e})}{Z}. \tag{5.3}$$

Including both of these probabilities over f requires a normalizing constant $Z = \sum_{\boldsymbol{f}} P(\boldsymbol{f})P(\boldsymbol{f}|\boldsymbol{e})$, which need not be calculated since it is implicit in the WFST when

token, which corresponds to the possibility of a word in $\boldsymbol{f}$ not aligning to any of the words in e (but rather a separate `null` token). We discuss this further in §5.3.3.

sampling or using Dijkstra's shortest path algorithm.

Finding the shortest path corresponds to determining the most likely source $\boldsymbol{f}$ and alignments $\boldsymbol{a}$. Since distinct alignment paths may have the same source $\boldsymbol{f}$, $\hat{\boldsymbol{f}}$ is most accurately found by marginalizing over the alignments:

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmax}} \sum_{\boldsymbol{a}} P(\boldsymbol{x}|\boldsymbol{f})P(\boldsymbol{f})P(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e}). \tag{5.4}$$

However, for computational reasons we simply approximate $\hat{\boldsymbol{f}}$ with the source side of the most likely path:

$$\hat{\boldsymbol{f}} \approx \underset{\boldsymbol{f}}{\operatorname{argmax}} \max_{\boldsymbol{a}} P(\boldsymbol{x}|\boldsymbol{f})P(\boldsymbol{f})P(\boldsymbol{f},\boldsymbol{a}|\boldsymbol{e}). \tag{5.5}$$

5.3.2 Learning Translation Model Parameters

We assume independent Dirichlet priors over each of the constituent distributions in the translation model T (one conditional probability distribution for each f). This is chosen due to conjugacy of the Dirichlet distribution with the categorical distribution over words. The conditional posterior can be expressed as:

$$P(f|e) = \frac{c_A(e,f) + \alpha P_{base}(f)}{c_A(e) + \alpha} \tag{5.6}$$

where $c_A(e,f)$ is a count of how many times f has aligned to e thus far in a corpus-wide set of alignments A between each f and an e in the same parallel sentence (with counts attributed to by the current alignment path $\boldsymbol{a}$ discounted); $c_A(e)$ is a count of how many times e has been aligned to; P_{base} is a uniform prior and α determines the emphasis on the base distribution.

We now turn to the task of determining parameters which will allow us to find the $\hat{\boldsymbol{f}}$ as discussed above. The approach we use involves blocked sampling of alignments A, with each utterance as a block. We perform sampling of these alignments proportionally to their probability given the data and our prior, in effect integrating over all parameter configurations T:

$$P(A,\mathcal{F}|\mathcal{S},\mathcal{E};\alpha,P_{base}) \propto \int_T P(A,\mathcal{F}|\mathcal{S},\mathcal{E},T)P(T;\alpha,P_{base})dT\,, \tag{5.7}$$

where $\mathcal{F}$ are the source-language sentences sampled from the word lattices $\mathcal{S}$, while $\mathcal{E}$ are the corresponding translations, and T represents a translation model that we assume prior information about and whose parameters we integrate over.[3]

The approach uses blocked Gibbs sampling where each block corresponds to one utterance and thus one of the composed WFSTs constrained on the observed translation. The sampling of paths through these composed WFSTs can be achieved using the method of *forward-filtering/backward-sampling* as described in (Scott 2002; Johnson *et al.* 2007; Neubig *et al.* 2012a). This method first computes forward probabilities in the same way the forward-backward algorithm for hidden Markov models does, before sampling edges from the end state proportionally to the product of the forward probability and the edge weight, using these forward probabilities to yield a path with probability proportional to the total probability of the edges in the path.[4]

After sampling a path consisting of lexical alignments, the counts of those lexical alignments are added to the cache used to calculate the Dirichlet posterior of e given f as per Equation 5.6 before the next WFST is composed and sampled from. When all WFSTs have been sampled from, we can repeat the sampling, first removing the counts in the cache attributed to the current WFST being sampled from, before adding counts from the new sample as per the sampling approach of Neubig *et al.* (2012a).

With repeated sampling of A the samples will converge to the posterior in Equation 5.7. Sampling sets of alignments n times, we use these alignment sets $A_1 \dots A_n$ to create a set of point estimates $T_1 \dots T_n$ for T. We then average these parameters to create a final expected T for the purposes of decoding using the approach of §5.3.1.

5.3.3 Variations on Parameter Formulation

In the previous formulation we discussed using $P(e|f)$ as the translation model parameters. Aside from the problem of deficiency, another problem with using $P(e|f)$

[3]Equation 5.7 assumes $\mathcal{S}, \mathcal{E} \perp T$.

[4]No Metropolis-Hastings rejection step was used. The WFST is thus not an exact implementation, but due to the short length of the utterances we argue it is not worth the effort nor time cost in sampling.

is that values of e with a higher marginal probability $P(e)$ tend to have a higher conditional probability $P(e|f)$. This problem makes itself especially clear when permitting null tokens on the English side, as it leads to degenerate alignments where most f_i end up aligning to the null token since it is present in every sentence.

The most obvious alternative formulation include use of $P(f|e)$ instead. Additionally, we can use normalizations of these probabilities. Notably, we propose and test an approach that uses $\frac{P(f|e)}{\sum_{f'} P(f'|e)}$ (both in sampling and decoding) where each f' is a token occurring in the original lattice. The denominator does not equal 1, since not every possible f' from the corpus will occur in any given lattice. The normalization term does not explicitly affect distinguishing between different source words in the WFST when sampling or decoding. However, it aids in aligning to the correct target word e by biasing towards alignments where f is most likely relative to its peers given e. Improving the alignments this way thus affects the translation model and, subsequently, the future paths chosen when sampling or decoding.

We also introduce a lattice weight hyperparameter λ. The contribution of original lattice probabilities from the acoustic model and language model against the translation model probabilities can be increased by simply multiplying the negative log probabilities by λ.

5.3.4 Experimental Evaluation

Experimental Setup

For the experiments we used the Fisher and CALLHOME Spanish–English Speech Translation Corpus (Post *et al.* 2013), which conveniently offers Spanish word lattices and crowdsourced English translations. The LDC human transcriptions (Wheatley 1996; Graff *et al.* 2010) are used as a gold standard against which to evaluate the ASR. Our preprocessing involved lowercasing all text, and removing punctuation from both the Spanish and English sides. We additionally removed from the corpus a small number of empty sentences and empty lattices.[5]

[5]Evaluating the 1-best output against the transcript, we find differences to WERs reported in Post *et al.* (2013). These were somewhat higher, likely accounted for by preprocessing differences.

Parameter type	1-best TM	Lattice TM	
		$\alpha, \lambda = 1$	$\alpha, \lambda =$ best
$P(e\|f)$	0.555	0.559	0.556
$P(f\|e)$	0.552	0.542	0.541
$\frac{P(e\|f)}{\sum_{e'} P(e'\|f)}$	0.568	0.574	0.570
$\frac{P(f\|e)}{\sum_{f'} P(f'\|e)}$	0.547	**0.539**	**0.539**

Table 5.9: Word error rates of different parameter variations when tuning on the CALLHOME training set. ASR 1-best accuracy is 0.569.

To evaluate how harnessing the English translations can improve use of the Spanish word lattices, we evaluate the word error rate of the chosen path through the composed WFST against the LDC transcriptions. We compare our approach, which we refer to as *Lattice TM*, with a similar method where the translation model is instead trained from 1-best paths from the lattice using GIZA++ (Och and Ney 2003), which we refer to as *1-best TM*. Note that while we are ultimately interested in calculating phoneme error rates, this experimentation uses a corpus of word lattices to explore how word-level translation modelling can improve word-based ASR. Phoneme error rate evaluation will follow in Section 5.4.

Tuning and Choice of Parameterization

We tuned two hyperparameters on the CALLHOME training set of approximately 14.5 hours (results from the Fisher set are thus those from which insight is most reliably attained): the lattice weight, λ, and the concentration parameter of the Dirichlet distributions, α. Tuning involved a simple grid search of λ and α over the values 0.5, 1, 2, and 4. We found no substantial improvements over the setting of 1 for both parameters. At $\lambda = 4$ the WER began to increase, although it was still significantly better than the ASR 1-best WER, with these improvements robust for all hyperparameter combinations evaluated. With no strong motivation to do otherwise, we left these parameters at $\lambda = 1$ and $\alpha = 1$ and evaluated on the test sets.

Method	Fisher	CALLHOME
ASR 1-best	0.355	0.586
1-best TM	0.343	0.576
Lattice TM	**0.335**	**0.562**

Table 5.10: Word error rates on the Fisher and CALLHOME test sets.

During tuning (Table 5.9), we also evaluated the parameterizations discussed in §5.3.3. The best parameterization, $\frac{P(f|e)}{\sum_{f'} P(f'|e)}$, was used for the subsequent evaluation. In tuning, permitting null alignments in the translation model WFST reduced scores for all parameter variations, most notably $P(e|f)$. The results presented in this chapter are based on a model that uses no null alignments.

Experimental Results

Table 5.10 shows the main results, across both of the test sets. The 1-best TM outperforms the ASR baseline, but underperforms Lattice TM on both test sets.

Figures 5.7 and 5.8 illustrate the change in performance when the method is run on a varying amount of training data, both as subsets and supersets of the training data. When training data is sufficiently limited, the TM trained on the 1-best path adversely affects performance, increasing the WER. In contrast, Lattice TM remains robust.

Each of the plots has a vertical rule labelled 'Test set.' The method is unsupervised, always taking an input lattice and written translation. Word error rates can be evaluated across all the data that is fed to the model. However, for a consistent point of comparison as the amount of input data varies, we use a fixed test set. To the left of the rule, when the amount of data is smaller than this test set, the training data is a subset of this test set. To the right of the rule, we run the model on a superset of the test set.

The example in Figure 5.4 (from the CALLHOME test set) epitomizes why the TM learnt from the lattice outperforms the TM learnt from the 1-best path. The

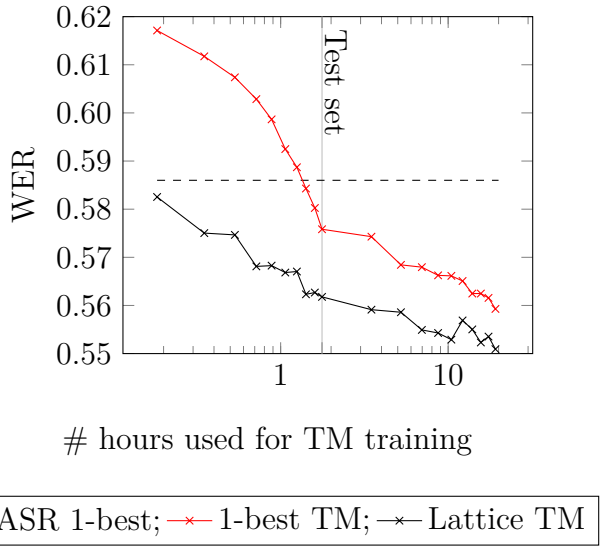


Figure 5.7: Speech recognition results on CALLHOME, which was the dataset used for tuning. We compare the performance of models trained on varying amounts of data. A fixed component of all the training data is used for word error rate evaluation, marked by the vertical line 'Test set' (since the method is unsupervised and designed to better transcribe data it is trained on, the training set and test set overlap).

crowdsourced English translation is *'aha, but one never knows'* and the gold transcription is *'Ahá, pero una nunca sabe.'*[6] The better path is the bottom one, choosing *'nunca'* over *'son.'* However, the 1-best path chooses *'son.'* Training a translation model from the erroneous 1-best path causes negative reinforcement, where the TM is even more likely to assign a high probability to *'son'* given *'never.'* Without enough training data to overcome this negative reinforcement, it has a substantial effect on scores.

Since *'never'* and *'nunca'* are relatively frequently occurring in the test data, the 1-best TM actually does assign a reasonable lexical translation score to this pair, however this is not enough to overcome the lattice's bias and the reasonable probability of *'son'* given *'never'* which was learnt from the erroneous transcription.

[6]The gold transcription is notably unobtainable from the lattice. Reduced pruning of lattices are likely to further improve scores.

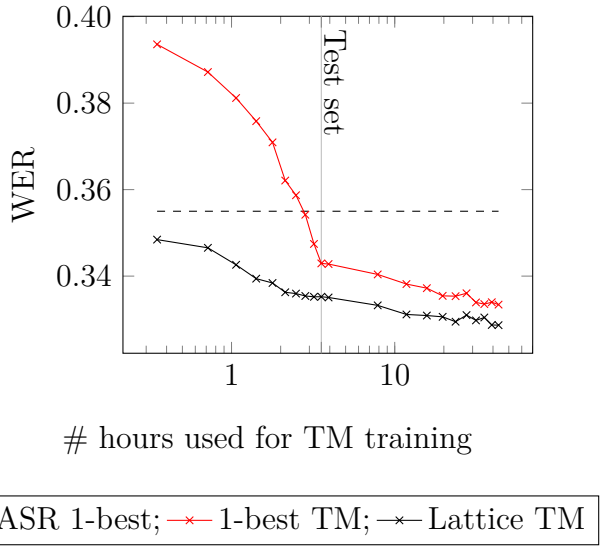


Figure 5.8: Speech recognition results on the Fisher corpus. We compare the performance of models trained on varying amounts of data. A fixed component of all the training data is used for word error rate evaluation, marked by the vertical line 'Test set' (since the method is unsupervised and designed to better transcribe data it is trained on, the training set and test set overlap).

Note that Lattice TM continues to outperform the 1-best TM approach when training data beyond the fixed test set is used. This suggests that Lattice TM gains an advantage from the information encoded in the lattice beyond avoiding the negative reinforcement of the 1-best TM approach.

Further examples from the Fisher test set are shown in Figure 5.11. Coloured words highlight how English reference words can inform how the Lattice TM approach deviates from the 1-best ASR path, or that of the 1-best TM. Sometimes this yields improvements with respect to the reference (green). Often the signal makes more sense or is justifiable (blue). Occasionally, it mistakenly biases away from the truth (red).

This method is very fast. Composition, sampling and caching for 1,000 utterances takes between 3 and 4 seconds on a single 1.80GHz Intel i7-4500U core. Running on the 213 minute Fisher evaluation set (Table 5.10) took less than 5 minutes, and scales

English:	i go because he doesnt dance either
Reference:	voy porque tampoco baila
1-best ASR:	pero por otro tampoco vaya
1-best TM:	sí pero tampoco vaya
Lattice TM:	sí pero tampoco bailar
English:	but viaja
Reference:	pero vi ajá
1-best ASR :	pero vi ajá
1-best TM:	pero vi ajá
Lattice TM:	pero que ajá
English:	and i live in athens mm
Reference:	y yo vivo en athens mh
1-best ASR:	y yo vivo en hacer
1-best TM:	y yo vivo en hacer
Lattice TM:	y yo vivo en athens
English:	i believe its something like christian science
Reference:	ah eh yo creo que christian science algo así
1-best ASR:	ah eh yo creo que kristin sáinz algo así
1-best TM:	ah ah ah eh yo creo que cristianos años algo así
Lattice TM:	ah eh yo creo que cristianos años algo así

Table 5.11: Examples from the Fisher test set (tokenized and lowercased). Each example consists of (top to bottom): crowd-sourced English translation (though code switching is common in both languages in the corpus); reference Spanish transcription; 1-best path through the word lattice; best path using a translation model trained on the 1-best paths (1-best TM); and best path using a translation model trained on lattices (Lattice TM). Green text indicates a useful correction based on learnt translations, along with blue text to a lesser extent. Red text indicates how a learnt translation pair can occasionally damage the transcription.

roughly linearly with more training data.

5.3.5 Implications

This experimentation demonstrated that having a written translation in another language can help improve speech recognition even when no pre-trained translation model is available. This is achieved by training a translation model directly on the ASR word lattices paired with the written translation in order to make the most of all information available in the lattice.

Although this is just a step towards a phoneme-based model, one natural setting for such an approach is for computer-aided translation of a small language for which there exists written data but no parallel corpora with the larger target language. However, since most languages have inadequate ASR technology, and stand to gain the most from improved speech recognition systems, future work should also strive to reverse the role of the languages in this setup, addressing the speech of a small language paired with a written translation in a larger language. Such bilingual data can be collected using a tool such as *Aikuma* (Bird *et al.* 2014b). However, for this to work, an ASR system with a lexicon and language model needs to be trained, perhaps using a tool such as *Woefzela* (De Vries *et al.* 2011). Otherwise this need should be sidestepped by working directly with the speech signal or phoneme lattices. We now discuss an extension to this model using phoneme lattices without a predetermined lexicon.

5.4 Learning a Lexicon and Translation Model from Phoneme Lattices

In the previous section we described how translation model parameters can be learnt from lattices paired with translations and demonstrated its effectiveness. However, the model assumes that a lexicon is available and that there are no out-of-vocabulary words in the speech recognition. In this section, we generalize the method to work instead with phoneme lattices, jointly learning the lexicon and translation

model in a Bayesian non-parametric framework, thereby harnessing translations to improve automatic phoneme recognition. The method assumes no prior lexicon or translation model, instead learning them from phoneme lattices paired with written translations of the speech being transcribed, assuming the target side is a major language that can be efficiently transcribed.

A Bayesian non-parametric model expressed with a weighted finite-state transducer (WFST) framework represents the joint distribution of source acoustic features, phonemes and latent source words given the target words. Sampling of alignments is used to learn source words and their target translations, which are then used to improve transcription of the source audio they were learnt from. Importantly, the model assumes no prior lexicon or translation model.

Experiments demonstrate that this method substantially reduces the phoneme error rate of transcriptions compared with a baseline recogniser and a similar model that harnesses only monolingual information, by up to 17% and 5% respectively. We also find that the model learns meaningful bilingual lexical items. The code is available online.[7]

Beyond the related work mentioned in §5.3, this method is related to various work discussed in Chapter 2, including work on phoneme translation modelling which has been done on 1-best transcriptions (Besacier *et al.* 2006; Stüker *et al.* 2009; Stahlberg *et al.* 2012; Stahlberg *et al.* 2014b), word segmentation in translation modelling (Chang *et al.* 2008; Dyer 2009; Nguyen *et al.* 2010; Chen and Xu 2015), and language model learning from lattices (Neubig *et al.* 2012a). Recent work that was done in parallel or after the contents of this chapter include translation modelling from speech directly, as described in Chapter 2.

5.4.1 Model Description

Our model extends the standard automatic speech recognition (ASR) problem and the word lattice approach of §5.3 by seeking the best phoneme transcription $\hat{\boldsymbol{\phi}}$ of an utterance in a joint probability distribution $P(\boldsymbol{x}, \boldsymbol{\phi}, \boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e})$ that incorporates

[7] `github.com/oadams/latticetm`

acoustic signal $\boldsymbol{x}$, phonemes $\boldsymbol{\phi}$, latent source words $\boldsymbol{f}$ and their alignments $\boldsymbol{a}$ to observed target transcriptions $\boldsymbol{e}$:

$$\hat{\boldsymbol{\phi}} = \operatorname*{argmax}_{\boldsymbol{\phi}} \sum_{\boldsymbol{f},\boldsymbol{a}} P(\boldsymbol{x}|\boldsymbol{\phi}) P(\boldsymbol{\phi}|\boldsymbol{f}) P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) \,, \tag{5.8}$$

assuming a Markov chain of conditional independence relationships. Note that bold symbols denote utterances as opposed to tokens. Deviating from standard ASR, no lexicon is given in training. Additionally, we replace language model probabilities with those of a translation model, also not given in training, and search for the best phoneme transcription instead of words.

Similar to Equation 5.4, the maximum approximation is used for tractability instead of summing over latent source words $\boldsymbol{f}$ that correspond to the phonemes $\boldsymbol{\phi}$:

$$\hat{\boldsymbol{\phi}} = \operatorname*{argmax}_{\boldsymbol{\phi}} \max_{\boldsymbol{f},\boldsymbol{a}} P(\boldsymbol{x}|\boldsymbol{\phi}) P(\boldsymbol{\phi}|\boldsymbol{f}) P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) \,, \tag{5.9}$$

Expression of the Distribution Using Finite-State Transducers

This model is distinct from the word-based model of the previous section since a prior lexicon is not assumed to be available. Instead, the lexicon must be learnt along with the translation model parameters.

We first describe how this model can be expressed in a WFST framework assuming we already have a lexicon and translation model, showing how this information can help find a more accurate phoneme transcription. We then describe how modifications can be used in order to jointly learn the lexicon and translation model.

We use a WFST framework to express the factors of Equation 5.9. The use of this framework is appealing for two key reasons. Firstly, it allows for computational tractability and simple inference via efficient methods FST composition and path sampling. Secondly, it allows for a modular, extendable framework, where components can be substituted with relative ease, facilitating simple future refinements to the model. Figure 5.9 uses a toy German–English error resolution example where an English translation guides the transcription of a German utterance to illustrate the components of the framework: a phoneme lattice representing phoneme uncertainty according to $P(\boldsymbol{x}|\boldsymbol{\phi})$; a lexicon that transduces phoneme substrings $\phi_{start}, \ldots, \phi_{end}$ of

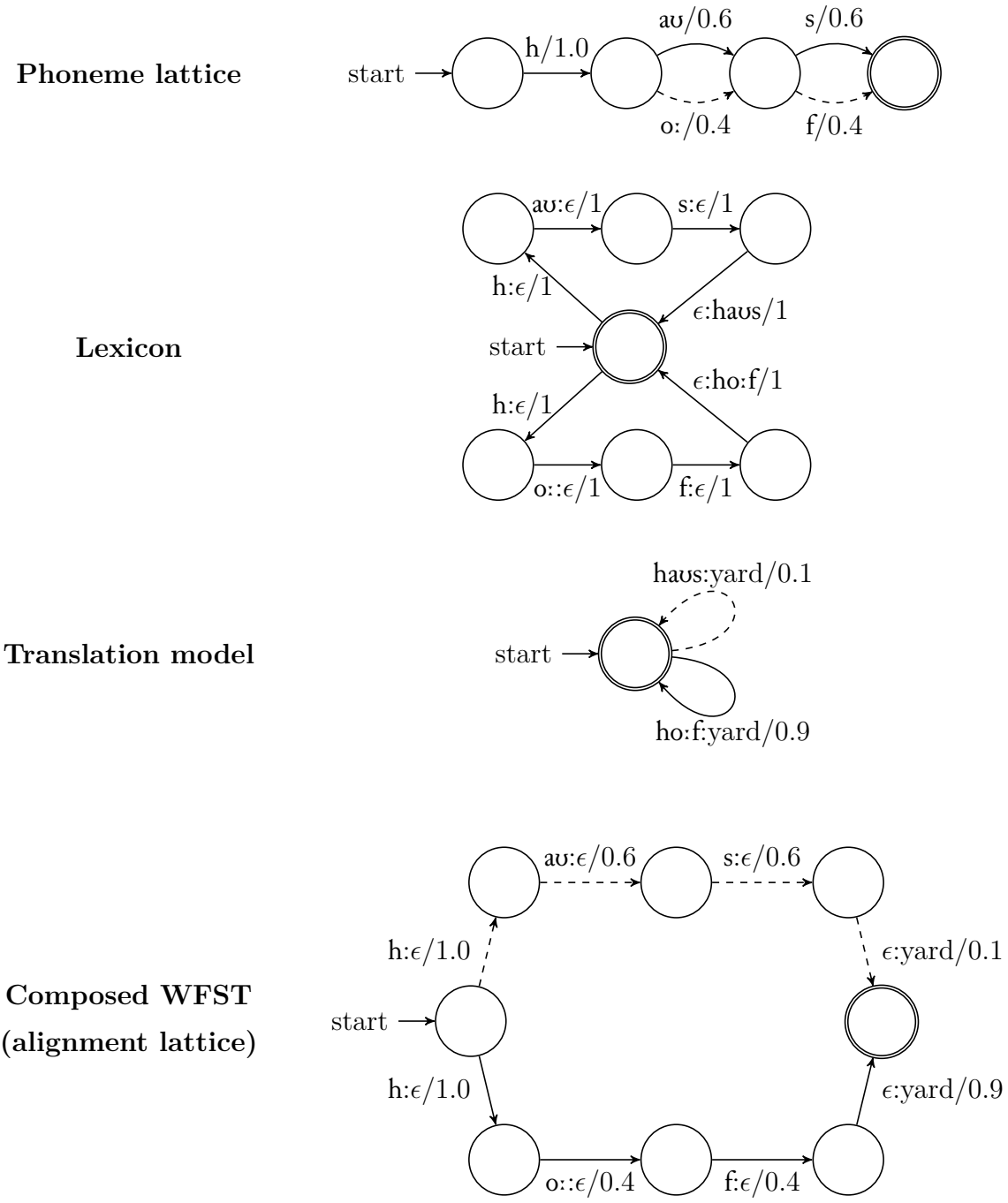


Figure 5.9: Top to bottom: the phoneme lattice, the lexicon, the translation model, and the resulting composed WFST. Given an English translation '*yard*,' the most likely transcription is corrected to /hoːf/ ('*Hof*') in the composed WFST, while in the original phoneme lattice it is /haʊs/ ('*Haus*'). Solid edges represent most likely paths.

$\boldsymbol{\phi}$ to source tokens f according to $P(\phi_{start}, \ldots, \phi_{end}|f)$; and a lexical translation model representing $P(f|e)$ for each e in the written translation. This is similar to the WFST architecture shown in Figure 5.3, except a separate lexicon FST is required to convert phoneme lattices into word lattices, which must be learnt during inference. The composition of these components is shown, illustrating how would-be transcription errors can be resolved. This framework is also reminiscent of the WFST framework used by Neubig *et al.* (2012a) for lexicon and language model learning from monolingual data.

We now describe how the factors of Equation 5.9 can be expressed in a weighted finite-state transducer framework. We describe each of the constituent components in turn. Note that though the transducer framework transduces from source phonemes to the target words, the generative story proceeds in the opposite direction: the observed text words produce the noisy acoustic signal.

Phoneme lattice The phoneme lattice captures phoneme sequences given by the acoustic model to explain the observed acoustic signal according to $P(\boldsymbol{x}|\boldsymbol{\phi})$.

Lexicon Each path represents one lexical entry and transduces from phonemes to a token that represents the sequence of those phonemes. The probability of each path is 1. This component captures $P(\phi_{start}, \ldots, \phi_{end}|f)$, where $\phi_{start}, \ldots, \phi_{end}$ represents a phoneme substring in the corpus that constitutes a word.

Translation model Each edge takes a foreign word token and yields an English word with probability $P(f|e)$. Since language model probabilities are not given as they were in §5.3, inclusion of this factor gives rise to the joint probability distribution of Equation 5.8, rather than the factored model of the previous section.

Composed FST The bottom of Figure 5.9, illustrates the composition of the lattice, lexicon and translation model. This *alignment lattice* expresses the distribution of Equation 5.9. Given this WFST, the 1-best path can be found by simply using Dijkstra's algorithm. Notice how the 1-best path through this path is informed by

all the components, and overcomes the uncertainty of the lattice alone, whose 1-best path would lead to the wrong transcription.

5.4.2 Learning the Lexicon and Translation Model

Because we do not have knowledge of the source language, we must learn the lexicon and translation model from the phoneme lattices and their written translations. In order to determine the translation model parameters as described above, we sample alignments A using the same approach described in §5.3.2. However, unlike that word lattice case, in this case the lexical entries need to be learnt. To do this, some adjustments must be made. The two most notable changes are that (a) the model must accommodate segmentation of phonemes into words not previously seen and (b) the translation model must become non-parametric, because the lexicon can be arbitrarily large. The lexicon needs to allow phonemes to pass through without conversion to known words, since they may constitute a word not yet in the lexicon. The extreme case of this is when the lexicon initially starts empty. For this we add an additional component (a word length model, described below) to the lexicon that serves as a prior distribution over sequences of phonemes.

We model lexical translation probabilities using a Dirichlet process. Let A be both the transcription of each source utterance $\mathbf{f}$ and its word alignments to the translation $\mathbf{e}$ that generated them. The conditional posterior can be expressed as:

$$P(f|e;A) = \frac{c_A(f,e) + \alpha P_{base}(f)}{c_A(e) + \alpha}\,, \tag{5.10}$$

where $c_A(f,e)$ is a count of how many times f has aligned to e in A and $c_A(e)$ is a count of e in any alignment; P_{base} is a base distribution that influences how phonemes are penalized based on their length; and α determines the emphasis on the base distribution. The distinction between Equation 5.10 and Equation 5.6 of the previous section is that the base distribution now takes a non-uniform, non-parametric form with probabilities varying with respect to the length of the word.

In order to express the Dirichlet process using the WFST components, we take the union of the lexicon with a word length model base distribution that consumes

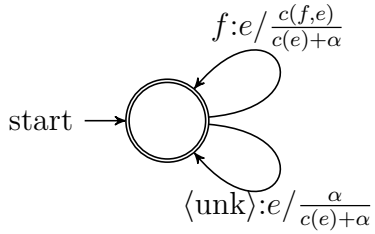


Figure 5.10: A translation model template. Each shown edge has a number of distinct instantiations for all $f \in V_f, e \in \boldsymbol{e}$.

a subsequence of phonemes $\phi_{start} \dots \phi_{end}$ and produces a special ⟨unk⟩ token with probability $P_{base}(\phi_{start} \dots \phi_{end})$. This ⟨unk⟩ token is consumed by a designated arc in the translation model WFST (see Figure 5.10 with probability $\frac{\alpha}{c_A(e)+\alpha}$, yielding a composed probability of $\frac{\alpha P_{base}(f)}{c_A(e)+\alpha}$. Other arcs in the translation model express the probability $\frac{c_A(f,e)}{c_A(e)+\alpha}$ of entries already in the lexicon. The sum of these two probabilities equates to Equation 5.10. The ⟨unk⟩ token is used when a new (unknown) token, not in the lexicon, is drawn from the base distribution. However, it can also be drawn from the base distribution even when the corresponding word is already in the lexicon.

See the translation model transducer as shown in Figure 5.10. The top edge corresponds to aligning a learned source word f to an observed target word e by drawing according to cached alignment counts. The bottom edge corresponds to the probability of drawing from the base distribution. There is only one ⟨unk⟩ token and the probability is the same given each $e \in \boldsymbol{e}$.

The word length model

The key distinction between the model in this section and that of §5.3 is this model's ability to segment the sequence of phonemes into words in order to perform unsupervised learning of a lexicon. Behind this segmentation is a base distribution —the word length model—which guides the segmentation. We experiment with three word length models, implementing WFSTs to represent them. These WFSTs can form part of the lexicon FST of Figure 5.9, by taking the union of the lexicon FST in that figure and the word length model WFST.

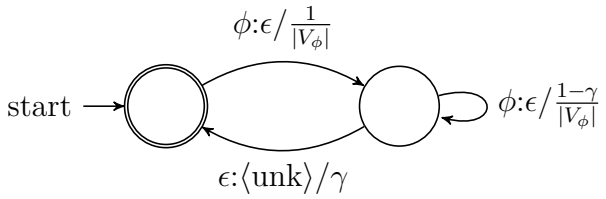


Figure 5.11: A word length model in the empty state that implements a Geometric prior, assigning probabilities to phoneme sequences not yet in the lexicon. Note that the arcs with phoneme ϕ on the input side can be considered templates, where in reality there is one such arc for each $\phi \in V_\phi$, where V_ϕ is the phoneme set.

These models can thus be considered simplified spelling models of (Mochihashi *et al.* 2009) and constitute the base distribution of the Dirichlet process.

Geometric distribution The most simple word length model is the geometric distribution, *Geometric(γ)*, shown in Figure 5.11. In this word length model prior (and the others discussed below in §5.4.2), all phoneme types are treated the same, with uniform probability. It is the length of the phoneme sequence that dictates its probability. In the case of the geometric distribution, longer sequences are always given less probability. This prior can most simply be implemented such that each phoneme type is given a uniform probability by this prior, with the probability of any length k phoneme sequence as:

$$P(k) = (1-\gamma)^{k-1}\gamma. \tag{5.11}$$

Disregarding the effects of other lexical items being in the lexicon, and of the weights of the supplied phoneme lattice, the most probable path is actually the one that segments the sequence as one large word. This is because there is a γ cost associated with completing a word. However, there are exponentially more segmentations of smaller words, and thus when sampling a path through the alignment lattice, smaller words are much more likely. Ultimately, with more sampling, these effects are outweighed by that of the translation model rewarding segmentations/alignments that explain the data.

The geometric prior described above is small, efficient, and helps to deliver phoneme recognition improvements in its own right. However, modelling word segmentation with geometric decay faces a key limitation in that the most likely length of a word is 1 phoneme, departing from observations of natural language.

Poisson distribution While the geometric distribution has a simple and elegant WFST formulation, it encourages the most common word length to be 1 phoneme, which doesn't encourage clustering of common groups of phonemes as much as we would like. A better prior would ensure the most likely length of a word is somewhat longer, but still with exponential decay thereafter. The Poisson distribution appears a more natural fit for this as a parameter λ is can be used to specify the average length. For any length k phoneme sequence, the probability is:

$$P(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{5.12}$$

While the geometric distribution can be easily expressed recursively, the Poisson distribution cannot. We limit the number of states in the Poisson WFST to 100. Furthermore, since the standard Poisson distribution would give non-zero probability to phoneme sequences of length 0, we also shift the Poisson distribution probabilities such that $P(k = 0;\lambda) = 0$ and the remaining probability mass takes the form:

$$P(k;\lambda) = \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!}. \tag{5.13}$$

Shifted geometric distribution The *shifted* geometric distribution, *Shifted(α, γ)*, like the Poisson distribution, mitigates the shortcoming of the geometric distribution whereby words of length 1 have the highest probability. It does so by having another parameter α that specifies the probability of a word of length 1, allowing it to be penalized, with the remaining probability mass distributed geometrically. An advantage of the shifted geometric distribution over the Poisson is that it admits a simple recursive WFST formulation, as shown in Figure 5.12, as well as the possibility of stronger penalization against single-phoneme words.

Figure 5.13 shows a histogram of the Poisson, Shifted, and Geometric distributions with the parameterizations found most effective during tuning.

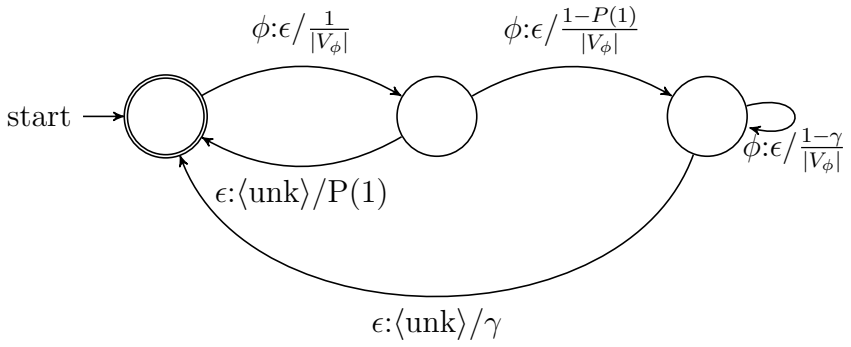


Figure 5.12: A shifted geometric distribution, with one probability specified, $P(1)$, giving the probability of a word being of length 1. Arcs with input ϕ are templates for all arcs that take a phoneme on the input. These arcs must have a denominator of $|V_\phi|$ since we uniformly distribute the probability over all phonemes in the vocabulary V_ϕ.

5.4.3 Experimental Evaluation

We evaluate the learnt lexicon and translation model by their ability to improve phoneme recognition, measuring phoneme error rate (PER).

Experimental Setup

We used up to 9 hours of English–Japanese data from the BTEC corpus (Takezawa *et al.* 2002), comprised of spoken utterances paired with textual translations. This allows us to assess the approach assuming quality acoustic models. We used acoustic models similar to Heck *et al.* (2015) to obtain source phoneme lattices without the help of a lexicon or language model. "Gold" phoneme transcriptions were obtained by transforming the text with pronunciation lexicons and, in the Japanese case, first segmenting the text into tokens using KyTea (Neubig *et al.* 2011a). Note that these transcriptions are good but not perfect. There may be some minor degree of phonemic deviation in the speech relative to what the pronunciation lexicon suggests about the pronunciation of each word in the lexicon. However, it makes sense to treat this phonemic transcription as a gold reference.

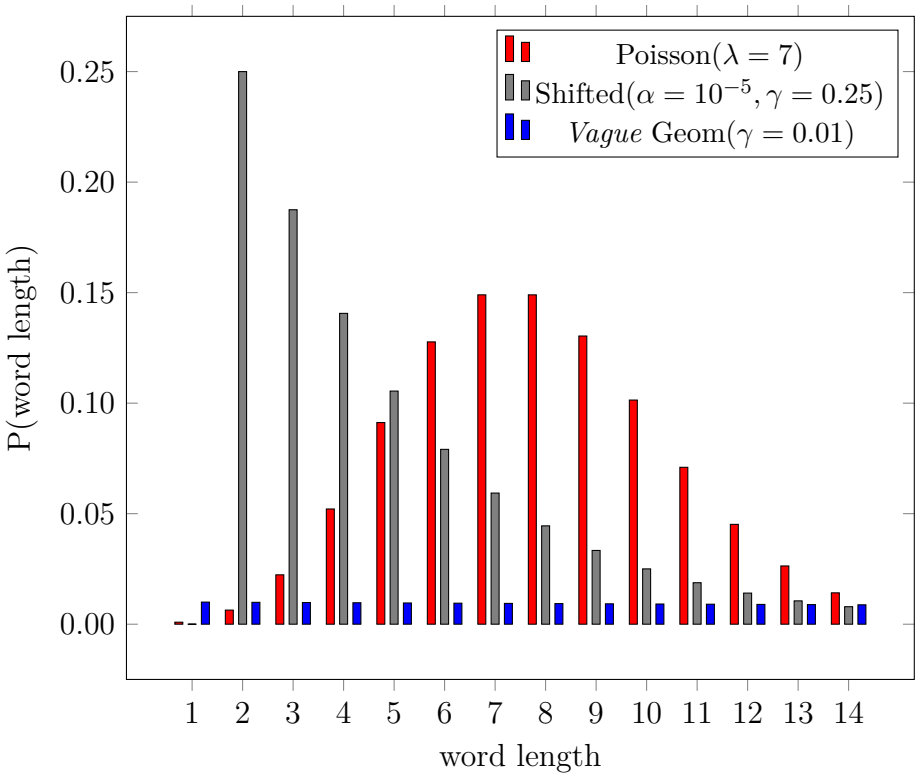


Figure 5.13: Histogram illustrating the word length model priors: the *Shifted* geometric distribution, a Poisson distribution, and a *Vague* geometric distribution.

We run experiments in both directions: transcription of English using Japanese translations (*en–ja*) and transcription of Japanese using English translations (*ja–en*), while comparing against three settings: the ASR 1-best path uninformed by the model (*ASR*); a monolingual version of our model that is identical except without conditioning on the target side (*Mono*); and the model applied using the source language sentence as the target (*Oracle*).

We tuned on the first 1,000 utterances (about 1 hour) of speech and trained on up to 9 hours of the remaining data.[8] Tuning was performed only using English in the monolingual configuration, using a grid search over α and the hyperparameters of the respective priors to find the best values. Only the oracle setup was used for tuning, with Geometric(0.01) (taking the form of a *vague* prior), Shifted(10^{-5}, 0.25)

[8]A 1 hour subset was used for PER evaluation.

	English (en)			Japanese (ja)		
	Mono	–ja	Oracle	Mono	–en	Oracle
ASR		22.1			24.3	
Vague	17.7	18.5	17.2	21.5	20.8	21.6
Shifted	17.4	16.9	**16.6**	21.2	**20.1**	20.2
Poisson	17.3	17.2	16.8	21.3	**20.1**	20.8

Table 5.12: Phoneme error rates (percent) when training on 9 hours of speech, averaged over 4 runs. As described in the text, *Mono* is a monolingual variation of the model that uses no translation, while *Oracle* uses a gold transcription in the same language as the phoneme recognition.

and Poisson(7) performing best.

We tuned α and γ. Changing α made little difference, though higher values began to reduce results, so we left it at 1. For γ, a value of 0.01 offered the largest PER reductions, with the PER appearing to monotonically decrease as γ decreased, from 0.213 to 0.196 against the baseline of 0.241.

Results

Table 5.12 shows en–ja and ja–en results for all methods with the full training data. Notably, English recognition gains less from using Japanese as the target side (en–ja) than the other way around, while the 'oracle' approach for Japanese recognition, which also uses Japanese as the target, actually underperforms ja–en. These observations suggest that using the Japanese target is less helpful, likely explained by the fine-grained morphological segmentation we used, making it harder for the model to relate source phonemes to target tokens.

While the Shifted and Poisson distributions were very close in performance, the vague geometric prior significantly underperforms the other priors. In the en–ja/vague case, the model actually underperforms its monolingual counterpart. The vague prior biases towards fine-grained English source segmentation, with words of length 1 most

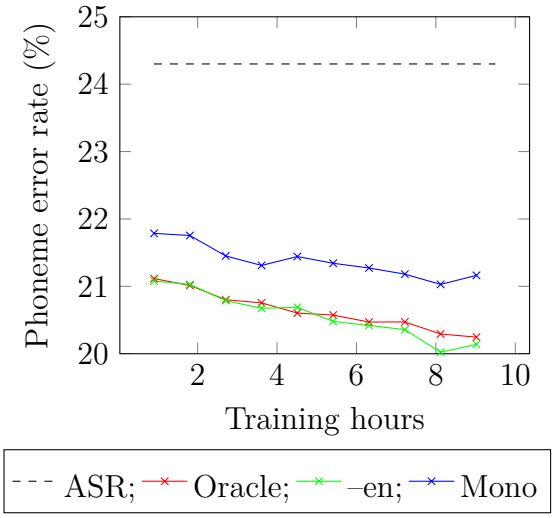


Figure 5.14: Japanese phoneme error rates using the *shifted* geometric prior when training data is scaled up from 1–9 hours, averaged over 3 runs.

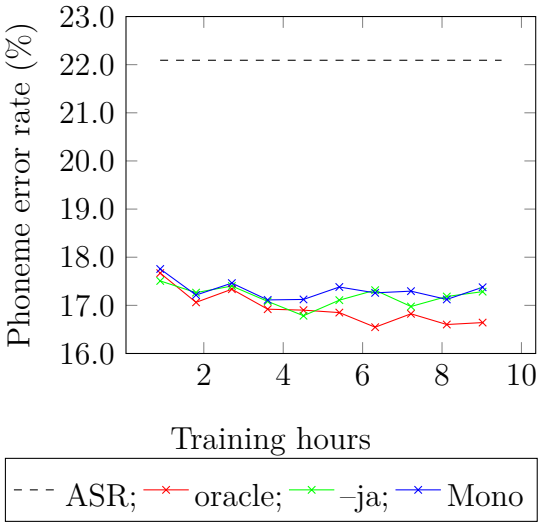


Figure 5.15: English phoneme error rates (expressed as percentages) of Poisson($\lambda = 7$) when training data is scaled up from 1 to 9 hours. A single run was used for each data point.

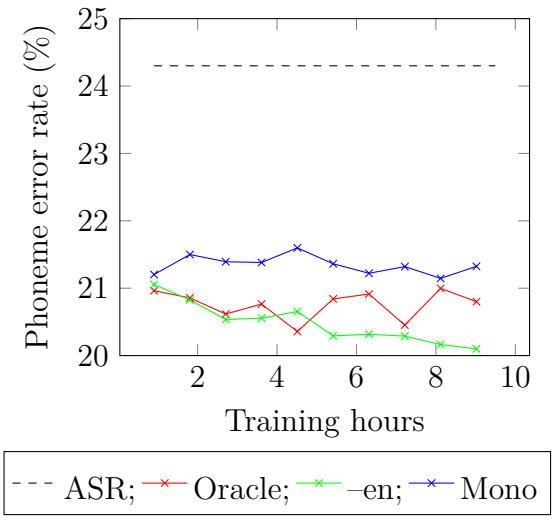


Figure 5.16: Japanese phoneme error rates (as percentages) using the Poisson prior when training data is scaled up from 1 hour to 9 hours. A single run was used for each data point.

common. In this case, fine-grained Japanese is also used as the target which results in most lexical entries arising from uninformative alignments between single English phonemes and Japanese syllables, such as /t/⇔ す. Perhaps the slight advantage Shifted gains over Poisson is because of its ability to even further penalize single-phoneme lexical items (see Figure 5.13), which regularly end up in all lexicons anyway due to their combinatorial advantage.

Figure 5.14 shows improvements of ja–en over both the ASR baseline and the Mono method as the training data increases, with translation modelling gaining an increasing advantage with more training data. To give a further sense of how performance generally scales with training data, Figures 5.15 and 5.16 present results of en–ja and ja–en respectively, using a Poisson prior. Since the method is unsupervised, the test set is a fixed 54 minute (1,000 utterances) subset of the training data.

While many bilingual lexical entries are correct, such as /wʌn/⇔ 一 (*'one'*), most are not. Some have segmentation errors /liːz/⇔ くださ (*'please'*); some are correctly segmented but misaligned to commonly co-occurring words /wʌt/⇔ 時 (*'what'* aligned to *'time'*); others do not constitute individual words, but morphemes aligned

Japanese:	この ブラウスは 思 ったより安かった
English:	this blouse was cheaper than i thought
Reference:	ð ɪ ʃ b l aʊ ʃ w ɔ z tʃ i p ɝ ð ʌ n aɪ θ ɔ t
1-best ASR:	ɪ ʃ p aʊ ʃ w ɪ θ tʃ i p ɝ ð aɪ θ ɔ d
Mono:	ɪ z b l aʊ ʃ w ʌ z tʃ i p ɝ ð ɛ n aɪ θ ɔ t
en–ja:	ð ɪ ʃ b l aʊ ʃ w ʌ z tʃ i p ɝ ð ɛ n aɪ θ ɔ t
Oracle (en–en):	ð ɪ ʃ b l æ ʃ i z tʃ i p ɝ ð ʌ n aɪ θ ɔ t
Japanese:	そう ですか
English	is that right
Reference:	ɪ z ð ʌ t ɹ aɪ t
1-best ASR:	ɪ z ð æ t ɹ eɪ
Mono:	ɪ z ð æ t ɹ aɪ t
en–ja:	ɪ z ð æ t ɹ aɪ t
Oracle (en–en):	ɪ z ð æ t ɹ aɪ t
Japanese:	小切手に 署名 しなければなりませんのね
English:	i have to sign the check dont i
Reference:	aɪ h æ V t u ʃ aɪ n ð i tʃ ɛ k d oʊ n t aɪ
1-best ASR:	h æ f t d ɪ ʃ aɪ d ɪ d tʃ ɛ k d ɑ n ʌ
Mono:	h æ f t ʌ ʃ aɪ n d tʃ ɛ k t ɑ n ʌ
en–ja:	h æ f t ʌ ʃ aɪ n d tʃ ɛ k t ɑ n ʌ
Oracle (en–en) :	aɪ h æ f t ʌ ʃ aɪ n d tʃ ɛ k t ɑ n ʌ

Table 5.13: Examples of output from four model variations. Top to bottom: Japanese translation; orthographic English transcription; phonemic English reference; shortest path through the ASR lattice (1-best ASR); monolingual method learning only from patterns in the lattice, without the translation or transcription (Mono); bilingual method harnessing the Japanese translation (en–ja); Oracle method harnessing the orthographic English transcription. Orange highlights the reference; green highlights corrections over the 1-best path; red highlights failures to correct; blue highlights a correction only the Oracle made on account of explicit information in the English transcription unavailable in the Japanese translation.

to common Japanese syllables /iːŋ/⇔ く (*'-ing'*); others still align multi-word units correctly /haʊmʌtʃ/⇔ いくら (*'how much'*). Note though that entries such as those listed above capture information that may nevertheless help to reduce phoneme transcription errors.

Table 5.13 shows some sentences from the corpus: the original English and Japanese transcriptions; the reference English phonemic pronunciation; the 1-best ASR output; the monolingual version of the model; the bilingual version that harnesses Japanese translations; and the Oracle, which harnesses orthographic English transcriptions.

The gold reference is shown in orange. In the first example, information from Japanese この is used to accurately transcribe /ð ɪ ʃ/ (*this*) in the en–ja method that harnesses the Japanese translations (shown in green). The Oracle similarly uses the orthographic English transcription *this* to the same end. In contrast, the 1-best ASR and Mono approaches fail to transcribe this word correctly (shown in red) since they do not have that information available. Similar instances are shown elsewhere in green, with the English and Japanese words for *though*, *right* and *sign* helping in their cases. Mono is often, but not always, able to learn monolingual word units for similar correction.

Blue text highlights a word only the Oracle method was able to correctly transcribe. This can be explained by the lack of explicit information on the Japanese side to learn that entry. A translation for *I* is not explicit in the Japanese text, thus en–ja fails to correctly transcribe /aɪ/. In contrast, the Oracle can resolve the error since it uses the orthographic English transcription, which has one-to-one correspondence with the speech.

5.5 Discussion

One of the appealing aspects of this final modular framework is that there is much room for extension and improvement. For example, by using adaptor grammars to encourage syllable segmentation (Johnson 2008), or incorporating language model probabilities in addition to our translation model probabilities (Neubig *et al.* 2012a).

The work presented in this chapter is consistent with the argument from the end

of Chapter 3 that bilingual lexical *coocurrences* can be useful for downstream tasks such as automatic speech recognition. This is true even when most entries from a simulated environment would not pass the litmus test of correctness based on whether they occur in a bilingual lexicon. However, bilingual context matters and can be very helpful. This holds true particularly for very different languages such as Japanese and English, where syntactic and morphological differences mean bilingual lexical items are even less likely to be correct in any strict sort of sense. However, despite this, significant reductions in phoneme error rate are found by harnessing the bilingual context, even when there is no prior information relating these languages available.

Despite drastic advances in acoustic modelling and speech recognition in the last decade, reducing phoneme error rates will remain important in low-resource contexts. In the case of the documentation of endangered languages, limited training data is available with which to train acoustic models and so making the most of available training data is key. Such data includes limited monolingual phonetic transcriptions, for which modelling must be effective. But importantly, such data may include other information, such as translations gathered in a language documentation setup based on Aikuma.

In this chapter we progressively removed simplifying assumptions until modelling speech and translations directly, assuming no prior lexicon, language model or translation model. However, the method did require an acoustic model with phoneme error rates between 20 and 25%. In a language documentation scenario, only limited supervised training data to train such acoustic models may be available. Future work might use a universal phoneme recognizer, making a step towards generalisability, or address acoustic modelling in the language with limited training data. The next chapter addresses the latter, exploring acoustic modelling for phonemic and tonal prediction in the context of language documentation for Yongning Na.

Chapter 6

Acoustic Modelling for Low-Resource Languages

Large portions of this chapter have appeared in the following papers:

> Oliver Adams, Trevor Cohn, Graham Neubig, Alexis Michaud (2017) Phonemic transcription of low-resource tonal languages, in *Proceedings of the Australasian Language Technology Association Workshop 2017 (ALTA)*, Brisbane, Australia. pp. 53–60.
>
> Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, Alexis Michaud (2018) Evaluating phonemic transcription of low-resource tonal languages for language documentation in *Proceedings of LREC 2018: 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan. (To appear).

The work so far in this thesis has assumed some sort of acoustic model is available with which to make initial hypotheses of phoneme transcriptions. In Chapters 3 and 4 the assumption was of an error-free single transcription hypothesis (possibly manually obtained). In chapter 5, this assumption was progressively relaxed, ultimately resulting in the use of probabilistic output from real acoustic models for Spanish, Japanese and English. Though this latter case is more realistic, it still assumed that ample data was available in these languages with which to train the acoustic models, which were subsequently used in simulated low-resource configurations. In the

final experimentation in Chapter 5, the phoneme error rates of the assumed acoustic models were in the order of 25%.

We now proceed to an exploration of acoustic modelling in low-resource contexts so that such lattices may be obtained. Another motivation for monolingual acoustic modelling comes from considering the workflow of the linguist documenting Yongning Na (Alexis Michaud), which involves phonetic transcription and translation occurring in parallel. While there is promise in harnessing translations to improve speech recognition, it is important to address the problem when information is only available in one language, which is a more general problem with broader applicability. In this chapter, we demonstrate that similar acoustic model performance that was assumed in the previous chapter can be obtained for languages with as little as 40 minutes of spontaneous narratives in Na, and as little as 25 minutes of elicited speech of Eastern Chatino, by using neural sequence-to-sequence models in a single-speaker context. Importantly, these languages are both tonal. Since tonal transcription is an important part of the language documentation workflow, we address tonal modelling too.

There has been work on low-resource speech recognition (Besacier *et al.* 2014), with approaches using cross-lingual information for better acoustic modelling (Burget *et al.* 2010; Vu *et al.* 2014; Xu *et al.* 2016; Müller *et al.* 2017) and language modelling (Xu and Fung 2013). However, speech recognition technology has largely been ineffective for endangered languages since traditional training pipelines, which generate orthographic transcriptions, require a large pronunciation lexicon and a language model trained on text. These speech recognition systems are usually trained on a variety of speakers and hundreds of hours of data (Hinton *et al.* 2012:92), with the goal of generalisation to new speakers. Since large amounts of text are used for language model training, such systems can rely on contextual information for tonal disambiguation via the language model (Le and Besacier 2009; Feng *et al.* 2012), and as a result often do not incorporate pitch information for speech recognition of tonal languages (Metze *et al.* 2013).

In contrast, language documentation contexts often have just a few speakers for model training, and little text for language model training. However, there may be

benefit even in a system that overfits to these speakers. If a *phonemic* recognition tool can provide a canvas transcription for manual correction and linguistic analysis, it may be possible to improve the leverage of linguists. The data collected in this semi-automated workflow can then be used as training data for further refinement of the acoustic model, leading to a snowball effect of better and faster transcription.

Note that in many language documentation scenarios there *will* be more information available than what we assume for the experimentation in this chapter, typically including a lexicon of some modest size as exploited in Chapter 4. In general, we advocate for the use of as much available information as possible (see Chapter 7), but in this chapter we use as training data only speech and phonemic transcriptions. Such experimentation establishes a lower bound on what can be achievable in these languages with the data available. A positive side-effect of not incorporating such information into the model is that it encourages the automatic transcription to be most faithful to the acoustic signal, since lexicons and language models bias towards prior information. This is desirable since the documentation work of Michaud has as a goal transcription of the speech with high phonetic accuracy, including fillers, fragments and deviations from canonical word forms.

In this chapter we investigate the application of neural speech recognition models to the task of phonemic and tonal transcription in a low-resource language documentation setting. We use the connectionist temporal classification (CTC) formulation (Graves *et al.* 2006) for the purposes of direct prediction of phonemes and tones given an acoustic signal, thus bypassing the need for a pronunciation lexicon, language model, and time alignments of phonemes in the training data. By sidestepping the requirement of a lexicon in this way, we make the use of automatic transcription technology more feasible in a language documentation setting. Moreover, by focusing on the single-speaker task, we set the models up to need less data.

We evaluate this approach on two tonal languages, Yongning Na and Eastern Chatino (Cruz and Woodbury 2006; Michaud 2017b). Na is a Sino-Tibetan language spoken in southwest China with three tonal levels, High (H), Mid (M) and Low (L) which can be combined for a total of seven tone labels. Eastern Chatino, spoken in Oaxaca, Mexico, has a richer tone set but both languages have extensive

morphotonology. Overall estimates of numbers of speakers for Chatino and Na are similar, standing at about 40,000 for both (Simons and Fennig 2017), but there is a high degree of dialect differentiation within the languages. The data used in the present study are from the Alawa dialect of Yongning Na, and the San Juan Quiahije dialect of Eastern Chatino. As a rule-of-thumb estimate, it is likely that the Na materials would be intelligible to a population of less than 10,000.[1]

Though a significant amount of Chatino speech has been transcribed (Chatino Language Documentation Project 2017), its rich tone system and opposing location on the globe make it a useful point of comparison for our explorations of Na, the language for which automatic transcription is our primary practical concern. Though Na has previously had speech recognition applied in a pilot study (Do *et al.* 2014a), phoneme error rates were not quantified and tone recognition was left as future work.

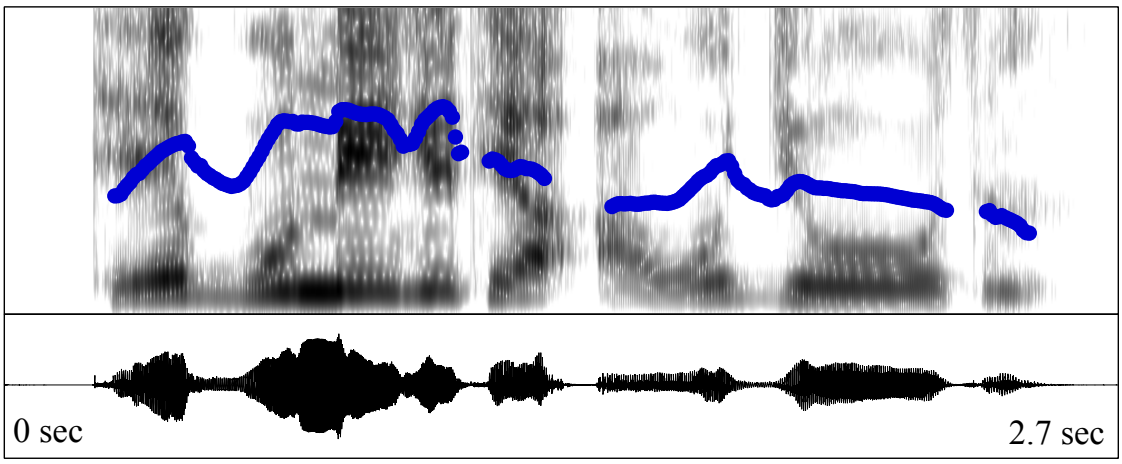
We perform experiments scaling the training data, comparing joint prediction of phonemes and tones with separate prediction, and assessing the influence of pitch information versus phonemic context on phonemic and tonal prediction in the CTC-based framework. Importantly, we qualitatively evaluate use of this automation in the linguist's transcription of Na. The effectiveness of the approach has resulted in its incorporation into the linguist's workflow. Our open-source implementation is available online.[2]

In summary we explore the effects of (a) scaling a training data to extremely low amounts; (b) exploring tonal prediction in the CTC framework; (c) analyzing errors on phoneme and tonal prediction with (d) a discussion of the outcomes of implementing this technology in a linguistic workflow.

6.1 Languages and Data

We now describe the Na and Chatino datasets in more detail.

[1]For details on the situation for Eastern Chatino, see Cruz (2011:18-23).
[2]`github.com/oadams/mam`



/tʰi˩˥, | go˧mi˧-dʑo˥ | tʰi˩˥, | ɑ˩ʁo˧ dʑo˩-tsɯ˩ | -mɣ˩ |/

As for the sister, she stayed at home.

Quant à la sœur, elle demeurait à la maison, dit-on.

而妹妹的话，留在家里。

tʰi˩˥	go˧mi˧	dʑo˩	tʰi˩˥	ɑ˩ʁo˧	dʑo˩	tsɯ˧˥	mɣ˩
-then	-little sister	-°top	-then	-home/family	-°exist.animated_beings	-°rep	-°affirm

Figure 6.1: A sentence from the Na corpus. Top to bottom: spectrogram with F_0 in blue; waveform; phonemic transcription; English, French and Chinese translations; morpheme-level English gloss.

6.1.1 Yongning Na (Mosuo, Narua)

Na is spoken by about 40,000 speakers (Yang *et al.* 2009) and is a *threatened* language: status 6b on Ethnologue (Lewis *et al.* 2015). It's also a richly tonal language (Yang *et al.* 2009), which bears more challenges for speech recognition. In particular, the morphotonology (tone sandhi) of the language is rich: the surface forms of tones for words depend on the surrounding context and deviate from their canonical form one might find in a lexicon. Additionally, the pitch of two syllables with the same tone varies depending on the context. This feature of the language exemplifies more general issues of prosody and sandhi found in many languages.

We use the Na corpus that is part of the Pangloss collection (Michailovsky *et al.* 2014). This corpus consists of around 100 spoken narratives from one speaker in the form of traditional stories, and spontaneous narratives about life, family and customs (Michaud 2017b:33). Currently the Na data consists of 14 hours of speech, 5.5 hours of which are accompanied by phonemic transcriptions. 200 minutes of this subset are spoken narratives segmented at approximately the sentence level, while 130 minutes are elicited speech of short utterances (Michaud 2017a). An example sentence entry can be seen in Figure 6.1.

We used up to 149 minutes of the transcribed spontaneous speech for training, 24 minutes for validation and 23 minutes for testing. The total number of phoneme and tone labels used for automatic transcription was 78 and 7 respectively.

In the phonemic transcription of the audio, the transcriptions also provide some markup to denote corrections and mistakes. Some phoneme sequences occur within angle brackets (< and >). These denote literal phoneme sequences that occurred as a mistake on the part of the speaker, such as gap-fillers and sentence fragments that were started and then respoken.

As a complement to these mistakes, phonemes occurring within square brackets ([and]), were omitted by the narrator, but were added in the process of transcription and translation in order to make a smoother translation.

6.1.2 Eastern Chatino

As a point of comparison for our primary work on Na, we evaluate the approach on another richly tonal language (with 14 tone classes, and 31 phoneme labels) from the other side of the world. Similar to Na, overall estimates for the number of speakers of Chatino stand at about 40,000 (Simons and Fennig 2017), with a high degree of dialect differentiation within the languages. Though a significant amount of Chatino speech has been elicited with corresponding transcriptions (Chatino Language Documentation Project 2017), we used data of Eastern Chatino dialect of San Juan Quiahije, Oaxaca, Mexico (spoken by approximately 3,000 speakers and rarely written) from the GORILLA language archive (Cavar *et al.* 2016) for the purpose of comparing

phoneme and tone prediction with Na when data restriction is in place.

This corpus was created with the idea of using speech recognition for more efficient language documentation in mind. Unlike the spontaneous speech of the Na corpus, this corpus is read speech of transcripts and texts. The recordings are free from background noise, variation in sound quality, and interaction between speakers mid recording (Cavar *et al.* 2016). These recordings were created with forced alignment in mind for their work.

6.2 Model

The underlying neural network used is a long short-term memory (LSTM) recurrent neural network (Hochreiter and Schmidhuber 1997) in a bidirectional configuration (Schuster and Paliwal 1997). The network is trained with the connectionist temporal classification (CTC) loss function (Graves *et al.* 2006). This is achieved through the use of a dynamic programming algorithm that efficiently sums over the probability of neural network output labels that correspond to the gold transcription sequence when repeated labels are collapsed and a special blank token is removed.

There have been advances in sequence-to-sequence modelling for speech recognition beyond the use of CTC, such as attentional neural architectures (Bahdanau *et al.* 2014; Chorowski *et al.* 2015; Bahdanau *et al.* 2016; Kim *et al.* 2016a), segmental conditional random fields and segmental neural networks (He and Fosler-Lussier 2012; O. Abdel-Hamid L. Deng and Jiang 2013; Lu *et al.* 2016). Though we acknowledge alternative models beyond CTC are applicable, we restrict the scope of this chapter to explore a comparison of varying training objectives and input features within the CTC paradigm.

The use of a recurrent neural network allows the model to implicitly model context via the parameters of the LSTM, despite the independent frame-wise label predictions of the CTC network. It is this feature of the architecture that makes it a promising tool for tonal prediction, since tonal information is suprasegmental, spanning many frames (Mortensen *et al.* 2016). Context beyond the immediate local signal is indispensable for tonal prediction, and long-ranging context is especially important in the

case of morphotonologically rich languages such as Na and Chatino.

We compare several training objectives for the purposes of phoneme and tone prediction. This includes separate prediction of 1) phonemes and 2) tones, as well as 3) jointly predicting phonemes and tones using one label set. Figure 6.3 an example of these three objectives using the sentence from the Na corpus shown in Figure 6.1. Since there is one tone per syllable in Na, and syllables can be deterministically segmented given phonemes, merging such independent phoneme and tone predictions is not a problem.

One might expect joint modelling to lead to lower phoneme error rates (PERs) and tone error rates (TERs), since more contextual information is given to the model in a form of multi-task learning. We experiment with a range of configurations to explore the effect phonemic and tonal context has on error rates.

6.2.1 Connectionist Temporal Classification

The key merit of the CTC formulation (Graves *et al.* 2006) is that there do not need to be alignments between the speech features $\boldsymbol{x} = x_1, \ldots, x_T$ and the transcribed labels of the training set $\boldsymbol{z} = z_1, \ldots, z_U$, where $U \leq T$.

The neural network output at each timestep t is an unnormalized probability distribution over the output labels (e.g. phonemes or tones) and a blank symbol, $\varnothing$. This output distribution is denoted $\boldsymbol{y}_t = y_t^0, y_t^1 \ldots, y_t^K$, where y_t^0 is the probability of $\varnothing$ and y_t^k is the probability of the kth label in the list of available labels. In order to arrive at a labelling from these output distributions, one simple and fast approach is best path decoding, where the most probable label is taken at each time step to form a path $\boldsymbol{\pi} = \pi_1, \ldots, \pi_T$. $\boldsymbol{\pi}$ can then be collapsed by first removing duplicate non-blank labels before removing blank labels. This is cheap, but is not guaranteed to find the most probable labelling. Finding the most probable labelling, which requires summing over values of $\boldsymbol{y}$ corresponding to all possible labellings, is computationally intractable.

In training the neural network parameters, the maximum likelihood objective is to maximize the probability of the training labels given the speech input, $p(\boldsymbol{z}|\boldsymbol{x})$.

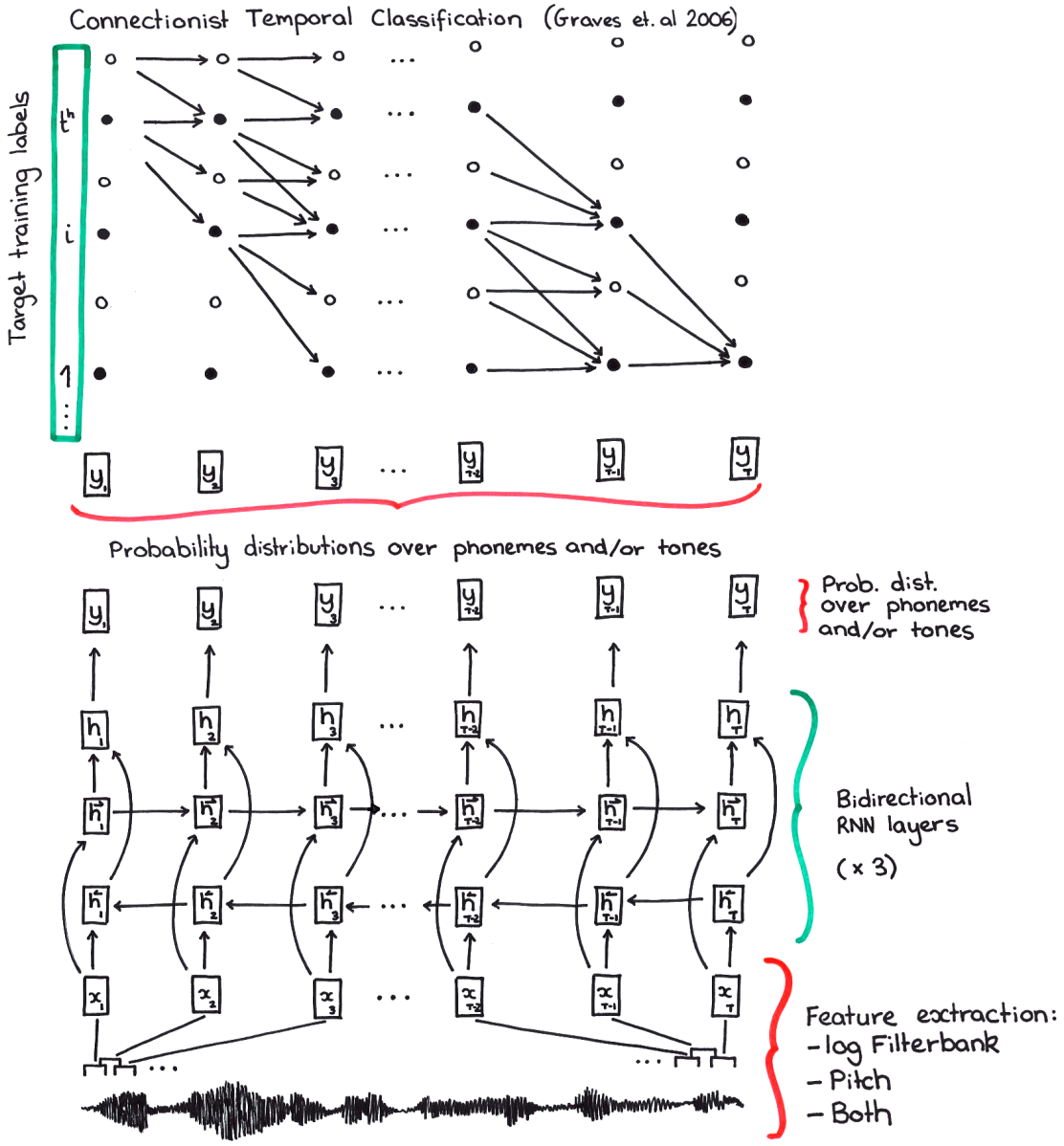


Figure 6.2: Neural network architecture with a CTC loss function. Dynamic programming allows for efficient computation of the probability of the relatively short label sequence given the input features $\boldsymbol{y}$, and for that probability to be expressed in a way differentiable with respect to any given timestep. Time alignment between the speech signal and phonemes is not required as prior knowledge. In the CTC graph, empty nodes represent blank symbols, while filled circles represent phonemes and tones. Note: only forward probabilities are illustrated, and we show one BiLSTM layer, though 3 are used in practice.

	Target label sequence
1. Phonemes	tʰ i g o m i dʑ o tʰ i ɑ ʁ o dʑ o t s ɯ m ɥ̩
2. Tones	˩˥ ˧ ˧ ˥ ˩˥ ˩ ˧ ˩ ˩ ˩
3. Joint	tʰ i ˩˥ g o ˧ m i ˧ d ʑ o ˥ tʰ i ˩˥ ɑ ˩ ʁ o ˧ dʑ o ˩ t s ɯ ˩ m ɥ̩ ˩

Figure 6.3: Target label sequences: 1. phonemes only, 2. tones only, 3. phonemes and tones together (*joint*).

Determining this probability requires summation over all of the paths $\boldsymbol{\pi}$ that collapse to $\boldsymbol{z}$ when duplicates and blanks are removed. This can efficiently be accomplished using dynamic programming to compute α and β probabilities similar to those used in forward-backward training of HMMs (Rabiner 1989). $\alpha_t^{z_u}$ is the probability of all CTC paths ending with label z_u at frame t, while $\beta_t^{z_u}$ is the probability of all CTC paths starting with z_u at frame t. The probability of the labelling given the speech features can then be expressed for any t as

$$p(\boldsymbol{z}|\boldsymbol{x}) = \sum_{u=1}^{2U+1} \alpha_t^{z_u} \beta_t^{z_u} ,$$

where we sum up to $2U + 1$ to account for blank labels being inserted between every non-blank label, allowing for blanks in output paths. This way of formulating $p(\boldsymbol{z}|\boldsymbol{x})$ allows it to be differentiated with respect to the neural network outputs at each timestep, facilitating updating of parameters via backpropagation.[3] An illustration is shown in Figure 6.2, where the CTC network sits on top of the LSTM.

6.3 Experimental Setup

We designed the experiments to answer these primary questions:

1. How do the error rates scale with respect to training data? What amount might linguists expect to transcribe before this technology can be used to aid in further transcription?

[3]For a more detailed explanation see Graves *et al.* (2006) and Miao *et al.* (2015).

2. How effective is tonal modelling in a CTC framework?

3. To what extent does phoneme context play a role in tone prediction?

4. Does joint prediction of phonemes and tones help minimize error rates?

We assess the performance of the systems as training data scales from 10 minutes to 150 minutes of a single Na speaker, and between 12 and 50 minutes for a single speaker of Chatino. Experimenting with this extremely limited training data gives us a sense of how much a linguist needs to transcribe before this technology can be profitably incorporated into their workflow.

We evaluate both the phoneme error rate (PER) and tone error rate (TER) of models based on the same neural architecture, but with varying input features and output objectives. Input features include log Filterbank features[4] (`fbank`), pitch features of Ghahremani *et al.* (2014) (`pitch`),[5] and a combination of both (`fbank+pitch`). These input features vary in the amount of acoustic information relevant to tonal modelling that they include. The output objectives correspond to those discussed in §6.2: tones only (`tone`), phonemes only (`phoneme`), or jointly modelling both (`joint`). We denote combinations of input features and target labellings as ⟨`input`⟩⇒⟨`output`⟩.

In case of tonal prediction we explore similar configurations to that of phoneme prediction, but with two additional points of comparison. The first is predicting tones given one-hot phoneme vectors (`phoneme`) of the gold phoneme transcription `phoneme`⇒`tone`). The second is the task of predicting tones directly from pitch features (`pitch`⇒`tone`). These important points of comparison serve to give us some understanding as to how much tonal information is being extracted directly from the acoustic signal versus the phoneme context.

Hyperparameters and Training

We used the LSTM cells proposed by Sak *et al.* (2014) in a bidirectional configuration. For feature extraction from the recordings we used 41 log Filterbank features

[4]41 log Filterbank features along with their first and second derivatives computed using the python_speech_features library on default settings

[5]As implemented in Kaldi

	Num. layers		
Hidden size	2	3	4
100	22.8	18.7	17.2
250	17.6	14.2	14.9
400	14.6	14.4	OOM

Table 6.1: Phoneme error rate (%) for `fbank`⇒`phoneme` on the Na validation set when trained on 2,048 utterances. *OOM* (out of memory) indicates that our system didn't have enough memory to complete experimentation with those hyperparameters.

along with their first and second derivatives. In some cases we also added pitch features of Ghahremani *et al.* (2014).

Our underlying model generally resembles that of Graves *et al.* (2013). For training, we used the Adam optimizer (Kingma and Ba 2015). We trained for a minimum of 40 epochs, thereafter stopping if no improvement was found for 10 consecutive epochs.

For tuning, we performed a small variety of grid searches to account for varying `input`⇒`output` configurations, in each case exploring 2,3 and 4 layers, and 100, 250 and 400 hidden units. We found 3 layers with 250 hidden units to be reasonably competitive in each `input`⇒`output` combination.

Training, Validation and Test sets

For both languages, preprocessing involved removing punctuation and any other symbols that are not phonemes or tones such as tone group delimiters and hyphens connecting syllables within words. We created training sets of varying sizes. For Na we used between 9.3 minutes and 149.8 minutes of transcribed speech, and for Chatino between 12 and 50 minutes. We also randomly created validation sets and test sets of sizes 24 and 23 minutes for Na, and 8 and 7 minutes for Chatino.

6.4 Quantitative Results

Figure 6.4 shows the phoneme and tone error rates for Na and Chatino.

Error rate scaling Error rates decrease logarithmically with training data. The best methods reliably have a lower than 30% PER with 30 minutes of training data. We believe it is reasonable to expect similar trends in other languages, with these results suggesting how much linguists might need to transcribe before semi-automation can become part of their workflow.

In the case of phoneme-only prediction, use of pitch information does help reduce the PER, which is consistent with previous work indication pitch information can aid in phoneme prediction (Metze *et al.* 2013).

Tonal modelling TER is always higher than PER for the same amount of training data, despite there being only 7 tone labels versus 78 phoneme labels in our Na experiment. This is true even when pitch features are present. However, it is unsurprising since the tones have overlapping pitch ranges, and can be realized with a different pitch over the course of a single sentence.

> "A consonant is much more context-independent than a tone: to cite just 1 factor, declination, the same tone (say, L) is realized on vastly different pitch at the outset of a sentence than at its end due to declination of F0 in the course of an affirmative sentence." —Alexis Michaud

This suggests that long-ranging context is more important for predicting tones than phonemes.

`fbank`⇒`tone` and `pitch`⇒`tone` are vastly inferior to other methods, all of which are privy to phonemic information via training labels or input. However, combining the fbank and pitch input features (`fbank+pitch`⇒`tone`) makes for the equal best performing approach for tonal prediction in Na at maximum training data. This indicates both that these features are complementary and that the model has learnt a representation useful for tonal prediction that is on par with explicit phonemic information.

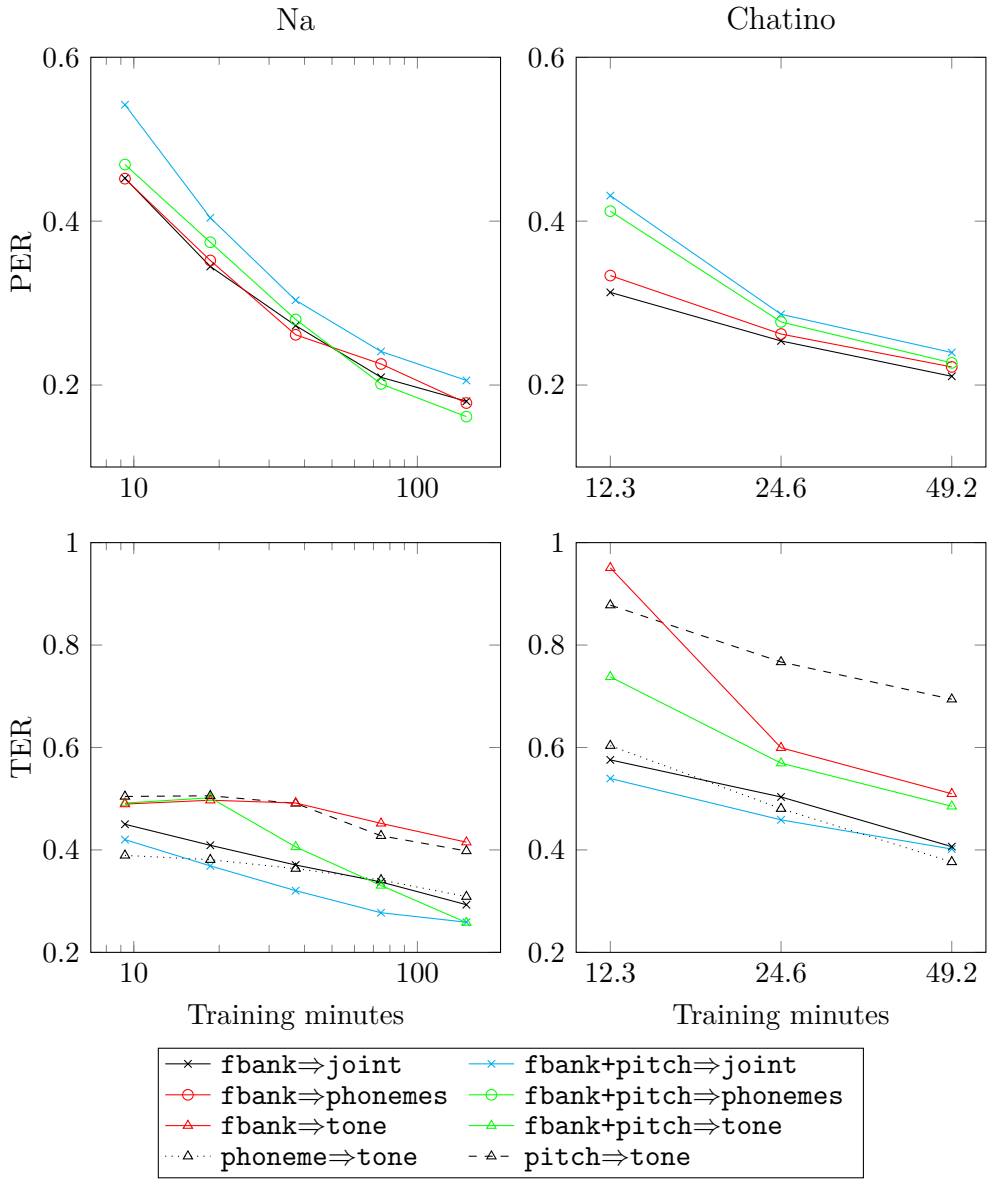


Figure 6.4: Phoneme error rate (PER) and tone error rate (TER) on test sets as training data is scaled for Na (left) and Chatino (right). The legend entries are formatted as ⟨input⟩ ⇒ ⟨output⟩ to indicate model input features and output labels.

Though tonal prediction is more challenging than phoneme prediction, these results suggest automatic tone transcription is feasible using this architecture, even without inclusion of explicit linguistic information such as constraints on valid tone sequences which is a promising line of future work.

Phoneme context To assess the importance of context in tone prediction, the `phoneme⇒tone` task gives us a point of comparison where no acoustic information is available at all. It performs reasonably well for Na, and competitively for Chatino. One likely reason for its solid performance is that long-range context is modelled more effectively by using phoneme input features, since there are vastly fewer phonemes per sentence than speech frames. The rich morphotonology of Na and Chatino means context is important in the realisation of tones, explaining why `phoneme⇒tone` can perform almost as well as methods using acoustic features.

Joint prediction Interestingly, joint prediction of phonemes and tones does not help phoneme prediction. In light of the celebrated successes of multitask learning in various domains (Collobert *et al.* 2011; Deng *et al.* 2013; Girshick 2015; Ramsundar *et al.* 2015; Ruder 2017), one might expect training with joint prediction of phonemes and tones to help, since it gives more relevant contextual information to the model. However, joint prediction of phonemes and tones does help with the tonal prediction. This makes sense since there are more phoneme labels than tone labels. As a result phonemes can provide more information to help tonal disambiguation than the other way around.

Na versus Chatino The trends observed in the experimentation on Chatino were largely consistent with those of Na, but with higher error rates owing to less training data and a larger tone label set. There are two differences with the Na results worth noting. One is that `phoneme⇒tone` is more competitive in the case of Chatino, suggesting that phoneme context plays a more important role in tonal prediction in Chatino. The second is that `fbank⇒tone` outperforms `pitch⇒tone`, and that adding pitch features to Filterbank features offers less benefit than in Na, also explained by

		Hypothesis				
		M	L	H	LH	MH
Reference	M	0	69.8	18.6	7.4	4.2
	L	77	0	14.6	6.1	2.3
	H	56.1	27.3	0	10.6	6.1
	LH	38.6	31.8	25	0	4.5
	MH	41.4	22.4	17.2	19	0

Figure 6.5: Confusion matrix showing the rates of substitution errors between tones (as a percentage, normalized per reference tone).

a greater influence of context.

Error Types

Figure 6.5 shows the most common tone substitution mistakes on the test set for the task of `fbank+pitch⇒joint`. Proportions were very similar for other methods. The most common tonal substitution errors were those between between M and L. Acoustically, M and L are neighbours; as mentioned above, in Na the same tone can be realised with a different pitch at different points in a sentence, leading to overlapping pitch ranges between these tones. Moreover, M and L tones were by far the most common tonal labels.

Table 6.2 shows an example automatically derived transcription and its reference from the test set.[6] Across the test set, the most common errors were substitution of L and M tones, both ways. Since MH occurs infrequently, the corpus is effectively imbalanced and hypotheses bias towards the more common L and M tones. It is also reasonably common that tones fail to be automatically transcribed. Every syllable in Na has a tone, and because of this fact future work should incorporate linguistic constraints to ensure hypotheses satisfy this rule. By enforcing valid tone sequences, it's likely many of these common errors would be prevented.

The second example in the table is from a variation of the task where only

[6]To hear the audio, see Sentence 224 at `http://lacito.vjf.cnrs.fr/pangloss/corpus/show_text_en.php?id=crdo-NRU_F4_FUNERAL_SOUND&idref=crdo-NRU_F4_FUNERAL`

Phoneme and tone prediction	
Reference	n j ɤ ˧ ɳ ɯ ˧ tʰ i ˧ tsʰ ɤ ˧˥ n i ˩ ʈʂʰ ɯ ˧ n j ɤ ˧ k i ˩ tʰ i ˧ tɕ ɯ ˧ tsʰ ɯ ˧˥
Hypothesis	n i ˧ ɳ ɯ pʰ i ˧ tsʰ ɤ ˩ n i ˩ ʈʂʰ ɯ ɻ̩ n j ɤ ˧ k i ˧ tʰ i ˧ tɕ ɯ ˩ tsʰ ɯ ˩

Phoneme-only prediction	
Reference	g ɣ w u h ĩ ʈʂʰ ɯ dʑ u tʰ i tsʰ e h õ pʰ u ɲ i z e m æ
Hypothesis	õ g ɣ i h ĩ ʈʂʰ ɯ dʑ u tʰ i tsʰ ɤ m õ h pʰ u ɲ i z e m æ

Table 6.2: Erroneous automatic transcription exemplifying typical errors. The top example is from the joint phoneme and tone prediction task. The bottom example is from the is phoneme-only prediction task. The reference transcription has punctuation, syllable boundaries and tone group delimiters removed.

phonemes are predicted. The automatic transcription includes a vowel at the start, which can actually be heard in the audio.[7] Exhaustive quantification of how frequent such phenomena are is costly and thus unknown, but some errors do stem from such fillers not part of the canonical word. A common error is the failure to transcribe voiced labial-velar approximants (/w/), which seem to be merged into the following (often erroneous) vowel predictions.

6.5 Qualitative Discussion

The phoneme error rates in the above quantitative analysis are promising, and are comparable to the performance of similar systems (Graves *et al.* 2013) on the TIMIT dataset (Garofolo *et al.* 1993). But is this system actually of practical use in a documentary linguistic workflow? We discuss here the experience of a linguist (Alexis Michaud) in applying this model to Na data to aid in transcription of 9 minutes and 30 seconds of speech. It is hard to quantify the ways in which incorporating automation

[7]To hear the audio, see Sentence 19 at `http://lacito.vjf.cnrs.fr/pangloss/corpus/show_text_en.php?id=crdo-NRU_F4_MOUNTAINS_SOUND&idref=crdo-NRU_F4_MOUNTAINS`

into the workflow affects the transcription process, but discussion of these effects is central to the present project.

We automatically transcribed recordings of two narratives for a total of nine minutes and 30 seconds of speech which had not previously been transcribed which were then used by the linguist as a starting point for manual correction and translation. Michaud subsequently published a blog post on his experience (Michaud 2017a), which was the basis for much subsequent discussion between us.

For the linguist, the process of revising the automatically generated transcription involves adding transcription elements not currently modelled by the system, such as punctuation and tone group breaks that delimit morphotonological processes (Michaud 2017a). In consultation with a native speaker, errors in the transcription are also corrected and comments are added, along with a sentence-level translation (initially in French, adding Chinese and English later as the need and opportunity arise).

Recognition Errors

The phonemic errors typically make linguistic sense: they are not random added noise. It was reported that they often bring to the linguist's attention phonetic facts that are easily overlooked because they are not phonemically contrastive.

One set of such errors is due to differences in articulation between different morphosyntactic classes:

> "For example, the noun 'person' /hĩ˥/ and the relativizer suffix /-hĩ˥/ are segmentally identical, but the latter is articulated much more weakly than the former and it is often recognized as /ĩ/ in automatic transcription, without an initial /h/. Likewise, in the demonstrative /ʈʂʰɯ˥/ the initial consonant /ʈʂʰ/ is often strongly hypo-articulated, resulting in its recognition as a fricative /ʂ/, /ʐ̩/, or /ʑ/ instead of an aspirated affricate. As a further example, the negation that is transcribed as /mõꜜ/ in *House-building2.290* instead of /mɤꜜ/. This highlights that the vowel in that syllable is probably nazalised, and acoustically unlike the average /ɤ/ vowel for lexical words. This is useful insight, but at the same time such cases illustrate a technical limitation of purely phoneme-based recognition over word-based speech recognition. The extent to which a word's

> morphosyntactic category influences the way it is pronounced is known to be language-specific (Brunelle *et al.* 2015); the phonemic transcription tool indirectly reveals that this influence is considerable in Na. Recognition 'errors' bring out specific patterns, demonstrating phonetic evolution and suggesting how grammatical morphemes follow phonetic evolutionary paths that are different from the rest of the lexicon." —Alexis Michaud

A second set is due to loanwords containing combinations of phonemes that are unattested in the training set. For example /ʑɯ˩pe˦/, from Mandarin *rìběn* (日本, 'Japan'). /pe/ is otherwise unattested in Na, which only has /pi/; accordingly, the syllable was identified as /pi/. In documenting Na, Mandarin loanwords were initially transcribed with Chinese characters, and thus cast aside from analyses, instead of confronting the issue of how different phonological systems coexist and interact in language use (Michaud 2017a).

A third set of errors made by the system result in an output that is not phonologically well formed, such as syllables without tones and sequences with consonant clusters such as /kgɣ̩/. These cases are easy for the linguist to identify and amend (Michaud 2017a).

In this sense the effective error rate is lower than the quantified phoneme error rate might suggest. Also noted is that the system copes well with overlong vowels, which an earlier Na speech recognition system had trouble with (Do *et al.* 2014a).

Tonal recognition remains an issue.

> "The recognition system currently makes tonal mistakes that are easy to correct on the basis of elementary phonological knowledge: it produces some impossible tone sequences such as M+L+M inside the same tone group." —Alexis Michaud

Very long-ranging tonal dependencies are not harnessed so well by the current model. This is consistent with quantitative indications in §6.4 and is a case for including a tonal language model or refining the neural architecture to better harness long-range contextual information.

Benefits for the Linguist

Using this automatic transcription as a starting point for manual correction was found to confer several benefits to the linguist.

Faithfulness to acoustic signal The model produces output that is faithful to the acoustic signal. In casual oral speech there are repetitions and hesitations that are sometimes overlooked by the transcribing linguist, who is engrossed in a holistic process involving interpretation, translation, annotation, and communication with the language consultant (Michaud 2017a). When using an automatically generated transcription as a canvas, there can be full confidence in the *linearity* of transcription, and more attention can be placed on linguistically meaningful dialogue with the language consultant.

Typographical errors and the transcriber's mindset Transcriptions are made during fieldwork with a language consultant and are difficult to correct down the line based only on auditory impression when the consultant is not available (Michaud 2017a). However, such typographic errors are common, with a large number of phoneme labels and significant use of combinations of keys (Shift, Alternative Graph, etc). By providing a high-accuracy first-pass automatic transcription, much of this manual data entry is entirely avoided. Enlisting the linguist solely for correction of errors also allows them to embrace a critical mindset, putting them in "proofreading mode", where focus can be entirely centred on assessing the correctness of the system output without the additional distracting burden of data entry.

Speed

> "Use of automatic transcription in fieldwork is only beginning. Assessing automatic transcription's influence on the speed of the overall language documentation process is beyond the scope of this paper and is left to future work. Language documentation is a holistic process. Beyond phonemic transcription, documentation of Na involves other work that happens in parallel: translating, discussing with a native, copying out new words into the Na dictionary, and being constantly on the lookout for new and unexpected linguistic phenomena." —Alexis Michaud

Further complicating this, the linguist's proficiency of the language and speed of transcription is dynamic, improving over time. This makes comparisons difficult (Michaud 2017a).

From this preliminary experiment, the efficiency of the linguist was perceived to be improved, but the benefits lie primarily in the advantages of providing a transcript faithful to the recording, and allowing the linguist to minimize manual entry, focusing on correction and enrichment of the transcribed document.

The snowball effect More data collection means more training data for better speech recognition performance. The process of improving the acoustic model by training on such semi-automatic transcriptions has begun, with the freshly transcribed *Housebuilding2* used in this investigation now available for subsequent Na acoustic modelling training. As a first example of output by incorporating automatic transcription into the Yongning Na documentation workflow, transcription of the recording *Housebuilding* was completed using automatic transcription as a canvas; this document is now available online (Michaud and Latami 2017) and is now being incorporated into further Na acoustic model training.

6.6 Summary

The experimentation and discussion of this Chapter addressed the task of phoneme and tone transcription in a resource-scarce context: that of a newly documented language. Beyond comparing the effects of various training inputs and objectives on the phoneme and tone error rates, we reported on the application of this method to linguistic documentation of Yongning Na. Its applicability as a first-pass transcription is very encouraging, and it has now been incorporated into the workflow.

These results give an idea of the amount of speech other linguists might aspire to transcribe in order to bootstrap this process: as little as 30 minutes in order to obtain a sub-30% phoneme error rate as a starting point, with further improvements to come as more data is transcribed in the semi-automated workflow.

We now proceed to the conclusions of this thesis, discussing the main findings and identifying promising future directions of work.

Chapter 7

Conclusion

The aim of the research has been to identify effective ways of using available data to establish a semi-automatic documentary linguistic workflow. Language documentation is currently very slow and there are not enough linguists documenting the world's languages to capture and build a record of them while they are still spoken. Automation or semi-automation of some parts of this work such as phonemic transcription can help that. Initial chapters explored using bilingual information to model the relationship between a low-resource source language and a high-resource target language. This included bilingual modelling of phonemic transcriptions and their translations as part of bilingual lexicon induction. Later, we modelled translated speech directly without the need for transcriptions. We argue that this has better potential as such translations of endangered language speech can be easier and quicker to collect than phonemic transcriptions (translations must be gathered in the end anyway, but have the potential to aid in automatic transcription and lexicon induction). We also explored the use of pre-existing bilingual lexicons and monolingual corpora to transfer information from high-resource languages to low-resource ones for improved language modelling, since language models are an important component of speech recognition and machine translation systems. Finally, the pressing concern of improving the use of technology in existing language documentation workflows motivated quantitative exploration of automatic phoneme transcription for tonal languages, along with a qualitative assessment of the incorporation of such techniques

into a linguist's workflow for the Yongning Na language of Southwestern China.

What can we conclude from this? We now relate the conclusions from the experiments conducted to the research questions set out in §1.2.1.

7.1 Main Findings

Translation models can learn hundreds of word and phrasal relationships with high precision from parallel text consisting of unsegmented phonemes and orthographic translations with as little as 1,000 parallel sentences Chapter 3 addressed research questions A1 and A2 stated in §1.2.1. It began with a preliminary experiment assessing machine translation performance, which partially addresses question A1, *Is translation modelling of unsegmented phonemic transcriptions effective?* Results show that machine translation performance, while substantially lower than that of word–word models, can yield BLEU scores over 17 with 241k training sentences, indicating that phoneme–word bilingual relationships have been captured between the languages. Competitive results were found between two very different phrase alignment approaches: that of traditional token alignment with heuristic phrase extraction, and that of joint alignment and segmentation in a hierarchical Bayesian framework.

For deeper insight into A1, and to answer A2 (*How do different translation models for this task compare in bilingual lexicon induction?*), we explored human evaluation of learnt phrase table entries across four different translation models. This intrinsic evaluation highlighted differences between translation models in the context of a task of importance in documentary linguistics: bilingual lexicon induction. In this context, we reduced the training data available drastically, to 10k parallel sentences. Such small quantities of data mimic the situation for many language documentation efforts, where data is very limited. Importantly, it makes the assumption of accurate phonemic transcriptions more reasonable, since these small amounts of data are feasible for trained linguists to collect and transcribe in a field linguistics scenario.

Findings show that approaches that have a statistical basis for clustering phonemes into word-like units perform better than standard machine translation heuristics. A

model that performs this statistical clustering jointly with alignment benefits more (PIALIGN) and is able to learn hundreds of bilingual lexicon entries with as few as 1,000 sentences of parallel phoneme–word data. PIALIGN along with unsupervised word segmentation and alignment had not been previously explored for this task, and outperformed other methods.

Bilingual lexicons can improve low-resource language modelling by enabling cross-lingual transfer of information from a large corpus in a high-resource language Chapter 4 first addresses Question C1 (*How can other bilingual resources, such as lexicons, be used to transfer information from a high-resource language to a low-resource language?*). We conducted an experiment to assess how well the quality of word embeddings in a low-resource language remain resilient in the face of limited data by harnessing distributional information from high-resource languages. Results on the English WordSim353 task, which assesses correlation of embedding similarity scores with human judgements, show large improvement over monolingual word embeddings trained on the same small amount of low-resource data when harnessing information across a variety of high-resource target languages. This also holds true when the language is of very dissimilar syntax or morphology to English, such as Japanese or Finnish.

Answering Question C1 via a method of learning cross-lingual word embeddings can be considered a prerequisite step before answering Question C2 (*Can such approaches be used to improve language modelling, which is useful to speech recognition and machine translation?*). We then deployed cross-lingual word embeddings, using them to initialize the parameters of neural network language models. Language models are an important component in machine translation and speech recognition systems, and are thus useful for semi-automatic language documentation. The approach offered consistent performance benefits in language modelling across a number of languages in low-resource simulations. However, application of this method to Yongning Na, a low-resource language spoken in Southwestern China, highlighted challenges in porting this technique to a language documentation setting. While a variety of factors likely played into this, two key issues were that many of the English

translations in the Na–English dictionary take non-standard forms and do not occur in the English Wikipedia corpus, and that there is a mismatch between the read and spoken speech (see §7.2). The morphotonology of Na presented another challenge, with canonical tones in the dictionary varying from the surface tones in the Na corpus. The findings suggest that these approaches may have more applicability in the low-resource languages for which more comprehensive dictionaries reflect the content of the monolingual documents in both source and target languages.

Translations of speech can help automatic phoneme transcription even when no prior information relating the languages is available. Chapter 5 addresses Questions B1 (*How can translation models be learnt from speech?*), and B2 (*Can these be used to improve speech recognition and automatic phonemic transcription?*), as well as addressing Question A1 in a context where the phoneme representation is errorful or uncertain, by using a lattice representation instead of the 1-best transcription.

Earlier experiments in the chapter highlighted fundamental issues with the notion of phoneme equivalence classes for modelling acoustic model errors. Most importantly, phonemes cannot be grouped neatly into classes that represent phonemes confusable with one another. Additionally, such an assumption is not easily amenable to the issue of errors in the form of insertion and deletion of phonemes, which are common in acoustic modelling.

These negative results and modelling failures prompted investigation into a more general and elegant model that takes as input phoneme lattices and their orthographic translations in order to jointly perform unsupervised word segmentation and alignment, learning a translation model and lexicon directly from phoneme lattices in order to improve phoneme transcription of those very lattices.

Empirical evidence suggested by a word-based variant demonstrated significant improvements in word-error rate, prompting further investigation of the full model, for which relative improvements of up to 17% over the original lattice were achieved. These results show sentence-level translations can be informative for phoneme recognition, even when no prior information relating the languages is available.

Phoneme and tonal transcription of a threatened language can produce transcripts of a high-enough quality to serve as a 'canvas,' aiding a linguist in their documentation workflow. Chapter 6 addresses Questions D1 and D2, which concern actual inclusion of such technology in a linguist's workflow. Such real-world evaluation is crucial: ultimately aiding language documentation is the aim of the technology, so evaluating it in that context is important. Equally important is that the concomitant dialogue with linguists that results from this evaluation is essential for guiding the future research direction, while granting insight into the language documentation process.

To this end, we address Question D1 (*How well can we predict tones for richly tonal languages?*), since transcription of transcription is important for many languages. We began with a quantitative experiment in phoneme and tone transcription by applying a neural speech recognition method in a low-resource context context: that of a newly documented language, Yongning Na. We compare the effects of various training inputs and objectives on the phoneme and tone error rate, which show that joint transcription of phonemes and tones allows for effective automatic tonal transcription in such a framework. Results on another language, Eastern Chatino, corroborate this finding. Beyond this quantitative evaluation, we address Question D2 (*How does phoneme and tonal prediction fit into the linguistic workflow?*), reporting on the application of this method to linguistic documentation of Yongning Na. Its applicability as a first-pass transcription is very encouraging, and it has now been incorporated into the workflow of Alexis Michaud, a linguist working on documenting and analyzing the language. Our results give an idea of the amount of speech other linguists might aspire to transcribe in order to bootstrap this process: as little as 30 minutes of clean recordings and transcriptions is needed in order to obtain a sub-30% phoneme error rate as a starting point, with further improvements to come as more data is transcribed in the semi-automated workflow.

7.2 Limitations and Future Work

7.2.1 Translation Modelling of Phonemes

Earlier work in Chapter 3 used phonemic transcriptions paired with translations. Such transcriptions can be considered a first approximation to the output of an automatic phoneme transcription tool with the generous assumption of no errors (created by converting orthographic text to phonemic representation, with accuracy limited to the capabilities of the text-to-speech system). Alternatively, when the amount of data is heavily restricted, as in §3.3, it becomes more reasonable to assume that a linguist has manually transcribed the speech with a high level of accuracy. However, throughout the course of the work it became clear that in such contexts the linguist will not blindly transcribe the phonemes, but will be conducting linguistic analysis of the language in conjunction with the transcription. Importantly, this will very likely involve creation of a lexicon and word segmentation of the transcription (along with glosses). These issues motivate the lattice-based approach of Chapter 5.

Furthermore, in the linguist's creation of a lexicon, much more than word token relationships are documented. While bilingual lexicon induction may be useful for highlighting such relationships a linguist has overlooked, the linguist's dictionary will include other content, such as explanations, examples, and part-of-speech tags.

7.2.2 Language Modelling

The cross-lingual language modelling work of Chapter 4 demonstrated reductions in perplexity for language models across a variety of languages in simulated low-resource settings. However, deploying the approach on Na yielded negative results, with no perplexity reductions. This reveals that the nature of the relationship between the dictionary entries and the content in the target monolingual corpora wasn't strong enough to harness cross-lingual distributional information, and suggests that in practice more preparation of the dictionary is required to reap the benefits that were seen when using the PanLex dictionary. This work is language-specific and has the potential to be very time consuming. Using source language corpora that are closer

in style to the target domain, such as using spoken narratives in the source language or perhaps conversational speech, would be an important next step.

Another limitation of the work is that only the intrinsic measurement of perplexity was used to evaluate the language models. An extrinsic evaluation of the method when the language model is integrated into a full speech recognition or machine translation pipeline would be more insightful and revealing of the value of such an approach. However, in extremely low-resource scenarios (for example, that of Yongning Na), a language model trained on a corpus of 2,000 sentences will have a high out-of-vocabulary rate when applied to new speech. In order for such a language model to be useful in speech recognition, incorporating character (or phoneme) level language modelling into a word-level language model has the potential to provide more accurate probability estimates for out-of-vocabulary words. Incorporation of such sub-word information at the level of characters or morphemes into language modelling has been explored (Lankinen *et al.* 2016; Ling *et al.* 2015; Verwimp *et al.* 2017; Shaik *et al.* 2011; Hao Fang *et al.* 2015), and future work improving on our approach should incorporate such information into a full speech recognition pipeline.

7.2.3 Translation Modelling of Speech

Earlier experiments in Chapter 5 use artificial German–English data, while subsequent models involve application to real Spanish–English and Japanese–English speech. While Japanese and English are very different languages, with strong results suggesting the methods ability to generalize, a key limitation of this work is that it is not applied to low-resource languages to improve language documentation.

A further limitation is that the acoustic model is trained in a supervised manner. While in Chapter 6 we show that such supervised training can yield comparably low phoneme error rates with less than an hour of transcribed speech, our approach is not incompatible with unsupervised or minimally-supervised approaches (Bansal *et al.* 2017b) and this should be explored. Moreover, further exploration of richer word length model priors would be valuable.

One could argue that there is a case for bypassing phonemic transcriptions of

endangered languages entirely. Phonemic transcription is part of the traditional documentary linguistics workflow and is useful for analysis and discussion. However, for the purposes of many speakers of unwritten languages, such transcription using Latin symbols can be perceived as a (relatively benign) form of neocolonialism. Some contemporaneous work to that of this thesis, discussed in Chapter 2, addresses spoken term discovery from the acoustic signal using bilingual information but without transcription. Maintaining sound as the primary representation of the language is useful for its speakers (in, say, the form of an audio dictionary), as it retains the most information. In directly modelling speech and translations, slight allophonic variations are not collapsed into a single symbol, while at the same time such approaches are not inimical to joint phonemic transcription for the linguist's purposes. Such approaches dealing with the source-language acoustic signal directly, along with translations, were not addressed in this thesis but constitute a promising line of future work.

7.2.4 Acoustic Modelling

The automatic method of phoneme transcription addressed in Chapter 6 is effective as part of a semi-automatic language documentation workflow. The scope for future work in this area is broad and promising, from back-end modelling improvements to user interface improvements which may allow for applicability of the method in more language documentation work.

Incorporating linguistic constraints There is much linguistic knowledge about Na that, if incorporated into the model, has the potential to reduce error rates. For example, we know that every syllable has a corresponding tone, yet sometimes syllables were transcribed without a tone. We also know consonant clusters are impossible, and that many tonal sequences, such as mid-low-mid, are impossible. Incorporating such hard constraints in decoding would be a simple first step. Alternatively, adjusting the model objective function in training to reflect these constraints may also be valuable. Beyond hard constraints, softer guidance in the form of a tonal language model that

better harnesses long-range dependencies than the frame-level recurrent neural network in the model may be valuable. Including coarser-grained lexical information as in standard speech recognition would be helpful, even when a comprehensive lexicon is not available (Liu *et al.* 2017b).

However, in doing this, there is a tradeoff between the faithfulness of the transcription to the acoustic signal, and the coherence of the transcription that must be considered (analogous to the machine translation faithfulness/fluency tradeoff). Linguists' goals may influence the relative priorities. One of the advantages of the approach was its faithfulness to the acoustic signal. In Chapter 6 we found that the model helped unearth some phonetic facts that might otherwise have been overlooked by the linguist. While including prior information into the model in the form of a language model may make for more coherent canonical transcriptions, this could be at the expense of losing some information in a narrow phonemic transcription useful for linguistic analysis.

The user interface The potential tradeoff between faithfulness and coherence of the transcription may be balanced through an appropriate user interface for the linguist. There is the potential for two transcriptions to be provided: one with minimal prior language knowledge for phonetic faithfulness, and another with stronger sources of prior information such as a word-level language model for coherence. The user would be able to choose between the two options, phoneme-by-phoneme or for stretches of the transcription.

The user interface has the potential to be enhanced in other ways. Automatically transcribed phonemes may be colour-coded by the model's confidence in the transcription, or the linguist could be presented with a pruned lattice to display other likely transcription possibilities. The stage is set for an interesting human–computer interaction problem which will need to be addressed.

In our work, the interface was crude: We applied the model to recordings of Na speech, and sent them to Michaud to use as a starting point in transcription. Better would be a piece of software useable to a less technical audience. Though existing software for training speech recognition models requires substantial technical

skills, this need not be the case. Linguists who are not computer scientists should be presented with an unintimidating graphical user interface where they can select audio files, phonemic transcriptions, supply a phoneme inventory and have a reasonable baseline model trained and tuned automatically. Such an interface could include the option to impose linguistic constraints, eg. the constraint that the consonant cluster /kg/ is not allowed.

Forced alignments Linguists are often interested in obtaining alignments at the phone-level between the speech and the transcription. A limitation of the CTC for this task is that it does not provide such alignment, since it sums over the many possible ways the speech signal might align to the phonemes. This is motivation for exploring other models, or exploring the use of secondary algorithms to extract time alignments from the CTC neural networks.

Language model incorporation While the recurrent neural network underlying the model implicitly learns a language model to capture context, a shortcoming of this CTC-based end-to-end modelling is that larger amounts of untranscribed texts cannot be leveraged in training. While not directly relevant to the Na dataset we use (all text has associated speech and is fed into the model), it is an open question of how to incorporate language model information into the CTC model, with language model information typically being incorporated separately to the end-to-end training (see Miao *et al.* (2015), for example).

Model architecture From the system architecture perspective, there is scope to explore various other neural architectures underlying the connectionist temporal classification (CTC) graph, such as convolutional neural networks (Zhang *et al.* 2016a; Wang *et al.* 2017b; Li and Wu 2016; Zhang *et al.* 2016a), their combination with recurrent neural networks (Li and Wu 2016), combining CTC with attention (Kim *et al.* 2016a) or segmental conditional random fields (Lu *et al.* 2017). We can also explore this task without CTC, using alternative models based on attention (Chorowski *et al.* 2015; Duong *et al.* 2016a; Bahdanau *et al.* 2016).

Transcribed speech is almost always accompanied by a much larger body of untranscribed speech. Semi-supervised speech recognition approaches that use untranscribed speech on top of transcribed speech can reach par performance with less training data (Dhaka and Salvi 2016). In light of the limited transcribed data, incorporating such semi-supervision into the model architecture is another obvious next step.

Broader deployment in other language documentation workflows for a richer understanding of limitations and necessary improvements Further exploration of such approaches in a diverse set of linguistic workflows is important to understanding how this technology can best be integrated into language documentation. Collaboration with Michaud revealed much information about how language documentation and analysis works in practice. This partnership between computer scientists and linguists is important, so that the practical needs of both parties can be met.

For many languages we won't have any transcribed speech at all for such supervised training. The experiments of Chapter 6 give an indication of how much transcribed audio needs to be acquired for meaningful performance. As it turns out, around an hour of speech from a single speaker yielded a 20% phoneme error rate, which is an acceptably low error rate for the purposes of post-editing.

This chapter exclusively modelled single-speaker speech recognition which is of value in language documentation settings. However, it is often desirable to create systems that generalize between speakers. In such a context, multilingual acoustic modelling, incorporating speech from diverse voices in other languages has the potential to aid in modelling speaker-dependent while capturing language-independent phonetic patterns.

7.3 Making the Most of What is Available

The best performing models for automatic phoneme transcription and bilingual lexicon induction would presumably make use of all the available information. Besides

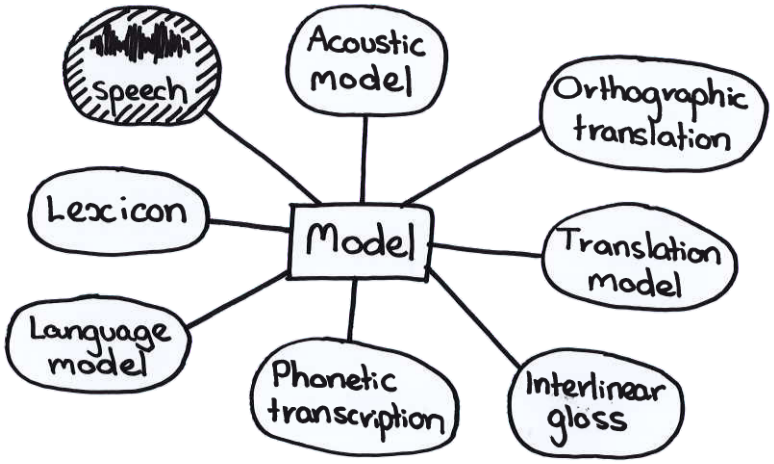


Figure 7.1: The ideal framework harnesses all available information. Speech is essential while the other variables are either observed or latent. The more observed variables there are, the better the inference for the latent variables.

speech, this information may include limited transcriptions, orthographic translations, interlinear glosses, bilingual lexicons, and, in the case of systems that need to generalize to multiple speakers, multilingual acoustic models or acoustic models from similar languages. Combining acoustic modelling capabilities with such information will allow for significant improvements in automatic phoneme transcription. However, we do not expect that automatic phoneme recognition alone will be error-free. Rather, combining automatic transcription with human expertise, as explored in Chapter 6, is likely the most efficient solution to the pernicious transcription bottleneck.

There has been work on models that facilitate more sophisticated inference to improve phoneme transcription. This includes work on monolingual models that learn a lexicon from speech (Neubig *et al.* 2012a; Lee *et al.* 2015) and the work presented in Chapter 5 that uses translations to learn lexicons and translation models. Bilingual information is also used in related work on phoneme-to-word alignment and lexicon induction (Stahlberg *et al.* 2012; Stahlberg *et al.* 2014b; Stahlberg *et al.* 2014a; Bansal *et al.* 2017a; Godard *et al.* 2016) and the work presented in Chapter 3, as well as speech-to-text alignment (Duong *et al.* 2016a; Anastasopoulos *et al.* 2016). Speech-

to-text alignment and lexicon induction are distinct tasks but related since models that jointly perform them are likely to enable better phoneme transcription.

The data required for these models are often a subset of a sparse but broad collection of information acquired in language documentation. While existing models that perform unsupervised inference of lexicons or that use more input data such as translations tend to perform better, an ideal framework shown in Figure 7.1 incorporates a larger array of possible available information for better statistical inference of the unavailable information.

Prior lexicons In many cases, a linguist investigating a language will have some pre-existing lexicon available to them. In the process of transcribing collected speech, the lexicon will be expanded. The models and experimentation described in Chapters 3, 5 and 6 assumed that no prior lexicon was available, but taking advantage of a prior dictionary has the potential to make more information available to the model with minimal extensions.

In Chapter 3 lexical entries a linguist is confident in can be incorporated into the model by simply appending confident phoneme–word entries as short parallel sentences to the end of the training data. This is model-agnostic. Alternatively, model-specific alternatives include adding explicit Pialign biparse trees for entries from an existing lexicon, which can be seamlessly integrated into inference.

In Chapter 5 prior lexical knowledge can be included by adjusting the model's cache counts based on our confidence. This works to a key strength of Bayesian models such as the one described in the chapter: the ability to incorporate prior belief about translations.

For the CTC-based neural network acoustic model of Chapter 6, incorporating lexical information is less straightforward. If the lexicon is extensive, then word-based decoding can be performed by composing a phoneme lattice with a lexicon WFST. In the more realistic case of a sparser lexicon, it is more difficult. One possible option is to use a mixed word/sub-word model where the lexicon biases towards likely words or morphemes without precluding the transcription of phoneme sequences not in the lexicon.

Multilingual acoustic modelling In the case where the language documentation requires acoustic model generalization across speakers, or where very limited supervised information is available, effective automatic phoneme transcription will likely depend on quality multilingual acoustic modelling. In such cases, the above frameworks would depend on using multilingual acoustic models, or acoustic models trained on phonetically similar languages, possibly adapted using a small amount of data from the target language. Though there is a history of work on cross-lingual and multilingual acoustic modelling (Köhler 1999; Schultz and Waibel 2001b; Vu *et al.* 2014; Imseng *et al.* 2014; Xu *et al.* 2016), there exist no widely available models for off-the-shelf application.

An ideal follow-on research project would simultaneously investigate multilingual acoustic modelling, models that can flexibly harness all information that may be available, as per Figure 7.1, and implementation of tools usable by linguists and speakers of the language that harness these models.

Data collection Another approach to best address the data sparsity issue consists in investigating approaches to guiding the data collection so that the process is most efficient. This may involve incorporating methods for inferring lexemes into the data collection process itself, offering hypotheses immediately to native speakers for their confirmation, rather than serving as a post-processing step. This sort of coupling of modelling and data collection may make more efficient use of the speakers' time while highlighting aspects of the lexicon that isn't well covered and allowing the models can guide the collection of data that is most informative in an active learning framework.

7.3.1 An Evaluation Suite for Methods on Bilingual Low-Resource Spoken Data

Automating language documentation tasks is difficult. Phoneme recognition is still out of reach for most languages, owing to the lack of transcribed data or a universal phone recogniser capable of generalizing to unseen languages. Although there has been research in the space of multilingual acoustic modelling (Köhler 1999;

Byrne *et al.* 2000; Schultz and Waibel 2001b; Heigold *et al.* 2013; Vu *et al.* 2014), there exists no widely available model with which to perform language-independent phone recognition or to facilitate easy adaptation thereof on a smaller in-domain dataset. In most cases, it is also not clear what sort of performance can be expected of such models. While there are many sources of bilingual data for low-resource languages, they are often not easily attainable and is often quarantined behind legal agreements. When it is accessible, the data is formatted in different ways and may require significant preprocessing to make them consistently amenable to a given machine learning method. Due to the lack of a communal dataset of speech and corresponding translations, procuring and preprocessing data must be done by each group of researchers, which results in different methods being applied to different datasets preprocessed differently. This makes comparison of methods difficult and unreliable. While corpora from the IARPA Babel project[1] offers data for a variety of low-resource languages, an important feature of the proposed dataset would include translations for training speech translation models as explored in Chapter 5.

We therefore recommend the development of a multilingual dataset along with a model so that baseline performance can be established for a number of languages. The availability of such a model, or the tools to train such a model, would facilitate research on downstream tasks such as speech-to-text alignment that depend on acoustic models but are hindered by the difficulty of the requirement of a full pipeline.

7.4 Towards a Research Program in Computational Documentary Linguistics

These activities (completed and projected) sit within a broader program of research to secure and leverage the full diversity of the world's languages. Creating a record of the world's languages will require advances in a) the rate of data acquisition, b) the modelling of available data and c) the useability of such software to enable its deployment by linguists who are doing the language documentation work. In addi-

[1] `www.iarpa.gov/index.php/research-programs/babel`

tion to helping documentation, improvements in data collection and the availability of language technology for threatened languages has the potential to aid communities of people speaking the languages who are interested in revitalization.

Most of this thesis constitutes exploration into modelling advances prompted by developments centering around the bilingual spoken data acquisition scenario of Aikuma (Hanke and Bird 2013; Bird *et al.* 2014b; Bird *et al.* 2014a; Hanke 2017). We have explored models designed to best make use of available data, and that can be connected with tools that enable the limited numbers of speakers to efficiently document their languages. Further refinement of these models, while creating user interfaces to make these methods broadly deployable in documentary linguistics is crucial. In light of rapid language extinction and language shift, this is a time-sensitive step that must be done while there is still time to capture and preserve the world's rich linguistic heritage.

Bibliography

Abney, Steven, and Steven Bird. 2010. The human language project: building a universal corpus of the world's languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 88–97.

Adda, Gilles, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: the BULB project. *Procedia Computer Science* 81.8–14.

Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: distributed word representations for multilingual NLP. *ArXiv:1307.16621*.

Alabau, Vicent, Luis Rodríguez-Ruiz, Alberto Sanchis, Pascual Martínez-Gómez, and Francisco Casacuberta. 2011. On multimodal interactive machine translation using speech recognition. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, 129–136.

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *ArXiv:1602.01925*.

Anastasopoulos, Antonios, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. 2017. Spoken term discovery for language documentation using translations. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, 53–58.

——, David Chiang, and Long Duong. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1255–1263.

Ataman, Duygu, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics* 108.331–342.

Audhkhasi, Kartik, Bhuvana Ramabhadran, George Saon, Michael Picheny, and David Nahamoo. 2017. Direct acoustics-to-word models for English conversational speech recognition. *ArXiv:1703.07754*.

Austin, Peter, and Julia Sallabank. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ArXiv:1409.0473*.

——, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4945–4949.

Bansal, Sameer, Herman Kamper, Sharon Goldwater, and Adam Lopez. 2017a. Weakly supervised spoken term discovery using cross-lingual side information. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 5760–5764.

——, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017b. Towards speech-to-text translation without speech recognition. *ArXiv:1702.03856*.

Barone, Antonio Valerio Miceli. 2016. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 121–126.

Bastan, Mohaddeseh, Shahram Khadivi, and Mohammad Mehdi Homayounpour. 2017. Neural machine translation on scarce-resource condition: a case-study on Persian-English. *ArXiv:1701.01854*.

Bellegarda, Jerome R. 2004. Statistical language model adaptation: review and perspectives. *Speech Communication* 42.93–108.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *The Journal of Machine Learning Research* 3.1137–1155.

——, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. *Proceedings of the 26th Annual International Conference on Machine Learning* 41–48.

Berard, Alexandre, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and translate: a proof of concept for end-to-end speech-to-text translation. *ArXiv:1612.01744*.

Bergmanis, Toms, and Sharon Goldwater. 2017. From segmentation to analyses: a probabilistic model for unsupervised morphology induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, 337–346.

Besacier, Laurent, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: a survey. *Speech Communication* 56.85–100.

——, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *Spoken Language Technology Workshop*, 222–225.

Bettinson, Mat, and Steven Bird. 2016. Developing a suite of mobile applications for collaborative language documentation. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, p. 156.

Bhatia, Parminder, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 490–500.

Bird, Steven. 1994. Automated tone transcription. In *Proceedings of the first meeting of the ACL special interest group on computational phonology (SIGPHON)*, 1–12.

——. 2011. Bootstrapping the language archive: New prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology* 6.1–16.

——. 2017. (To appear) Designing mobile applications for documenting endangered languages. In *Oxford Handbook of Endangered Languages*, ed. by Ken Regh and Lyle Campbell. Oxford University Press.

——, and David Chiang. 2012. Machine translation for language preservation. In *24th International Conference on Computational Linguistics*, p. 125.

——, Lauren Gawne, Katie Gelbart, and Isaac McAlister. 2014a. Collecting bilingual audio in remote indigenous communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1015–1024.

——, Florian R Hanke, Oliver Adams, and Haejoong Lee. 2014b. Aikuma: a mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 1–5.

Blachon, David, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the Lig-Aikuma mobile device app. *Procedia Computer Science* 81.61–66.

Blunsom, Phil, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A Gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, 782–790.

Bohac, Marek, Michaela Kucharova, Zoraida Callejas, Jan Nouza, and Petr Červa. 2014. A cross-lingual adaptation approach for rapid development of speech recognizers for learning disabled users. *EURASIP Journal on Audio, Speech, and Music Processing* 2014.39.

Botha, Jan A, and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, 1899–1907.

Brown, Peter E, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics* 19.263–312.

Brown, Peter F, Stanley F Chen, Stephen A Della Pietra, Vincent J Della Pietra, Andrew S Kehler, and Robert L Mercer. 1994. Automatic speech recognition in machine-aided translation. *Computer Speech & Language* 8.177–187.

Brunelle, Marc, Daryl Chow, and Thụy Nhã Uyên Nguyễn. 2015. Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. In *Proceedings of 18th International Congress of Phonetic Sciences*, 1–5.

Burget, Lukáš, Petr Schwarz, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek, Nagendra Goel, Martin

Karafiát, Daniel Povey, and Others. 2010. Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In *Acoustics Speech and Signal Processing, 2010 IEEE International Conference on*, 4334–4337.

Byrne, William, Peter Beyerlein, Juan M Huerta, Sanjeev Khudanpur, Bhaskara Marthi, John Morgan, Nino Peterek, Joe Picone, Dimitra Vergyri, and W Wang. 2000. Towards language independent acoustic modeling. In *Acoustics, Speech, and Signal Processing. IEEE International Conference on*, volume 2, 1029–1032.

Cai, Deng, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for Chinese. *ArXiv:1704.07047*.

Camacho-Collados, Jose, Ignacio Iacobacci, Roberto Navigli, and Mohammad Taher Pilehvar. 2016. Semantic representations of word senses and concepts. *ArXiv:1608.00841v1*.

Carl, Michael, Maite Melero, Toni Badia, Vincent Vandeghinste, Peter Dirix, Ineke Schuurman, Stella Markantonatou, Sokratis Sofianopoulos, Marina Vassiliou, and Olga Yannoutsou. 2008. METIS-II: low resource machine translation. *Machine Translation* 22.67–99.

Casacuberta, F., H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. García-Varea, D. Llorens, C. Martínez, S. Molau, F. Nevado, M. Pastor, D. Picó, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language* 18.25–47.

Caseli, Helena M, V Nunes das Graças, and Mikel L Forcada. 2006. Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. *Machine Translation* 20.227–245.

Cavar, Malgorzata E, Damir Cavar, and Hilaria Cruz. 2016. Endangered language documentation: bootstrapping a Chatino speech corpus. In *Proceedings of LREC 2016*, 4004–4011.

Chahuneau, Victor, Eva Schlinger, Noah a. Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* 1677–1687.

Chan, William, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. 2015. Listen, attend and spell. *ArXiv:1508.01211*.

Chandar AP, Sarath, Stanislas Lauly, Hugo Larochelle, Mitesh M. Khapra, Balaraman Ravindran, Vikas Raykar, and Amrita Saha. 2014. An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems 27*, 1853–1861.

Chang, Pi-Chuan, Michel Galley, and Christopher D Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, 224–232.

Chatino Language Documentation Project, 2017. Chatino Language Documentation Project Collection. The Archive of the Indigenous Languages of Latin America: www.ailla.utexas.org.

Chen, Stanley F, and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13.359–394.

Chen, Wei, and Bo Xu. 2015. Semi-supervised Chinese Word segmentation based on bilingual information. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1207–1216.

Chen, Welin, David Grangier, and Michael Auli. 2015. Strategies for training large vocabulary neural language models. *ArXiv:1512.04906*.

Chen, Yanqing, Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2013. The expressive power of word embeddings. *ArXiv:1301.3226*.

Cherry, Colin, and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation*, 17–24.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 263–270.

——. 2007. Hierarchical phrase-based translation. *Computational Linguistics* 33.201–228.

Cho, Kyunghyun, Bart Van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: encoder–decoder approaches. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* 103–111.

Chorowski, Jan, and Navdeep Jaitly. 2016. Towards better decoding and language model integration in sequence to sequence models. *ArXiv:1612.02695*.

Chorowski, Jan K, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in Neural Information Processing Systems 28* 577–585.

Chung, Junyoung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *ArXiv:1603.06147*.

Cieri, Christopher, and Mark Liberman. 2006. More data and tools for more languages and research areas: a progress report on LDC activities. In *5th International Conference on Language Resources and Evaluation*.

Clifton, Ann, and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 32–42.

Cohn, Trevor, and Gholamreza Haffari. 2013. An infinite hierarchical Bayesian model of phrasal translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 780–790.

Collobert, Ronan, Christian Puhrsch, and Gabriel Synnaeve. 2016. Wav2Letter: an end-to-end convnet-based speech recognition system. *ArXiv:1609.03193*.

——, and Jason Weston. 2008. A unified architecture for natural language processing. *Proceedings of the 25th international conference on Machine learning* 160–167.

——, ——, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12.2493–2537.

Conneau, Alexis, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *ArXiv:1710.04087*.

Costa-jussà, Marta R, and José A R Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Coulmance, Jocelyn, Jean-Marc Marty, Guillaume Wenzek, and Amine Benhalloum. 2015. Trans-gram, fast cross-lingual word-embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1109–1113.

Crowley, Terry. 2007. *Field linguistics: a beginner's guide.* Oxford University Press.

Cruz, Emiliana, 2011. *Phonology, tone and the functions of tone in San Juan Quiahije Chatino.* University of Texas at Austin dissertation.

——, and Tony Woodbury. 2006. El sandhi de los tonos en el Chatino de Quiahije. In *Las memorias del Congreso de Idiomas Indígenas de Latinoamérica-II*.

De Mulder, Wim, Steven Bethard, and Marie-Francine Moens. 2015. A survey on the application of recurrent neural networks to statistical language modeling. *Computer Speech & Language* 30.61–98.

De Vries, Nic J, Jaco Badenhorst, Marelie H Davel, Etienne Barnard, and Alta De Waal. 2011. Woefzela - an open-source platform for ASR data collection in the developing world. In *Proceedings of the Annual Conference of the International Speech Communication Association, (INTERSPEECH)*.

——, Marelie H Davel, Jaco Badenhorst, Willem D Basson, Febe De Wet, Etienne Barnard, and Alta De Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech communication* 56.119–131.

DeNero, John, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 314–323.

Deng, Li, Geoffrey Hinton, and Brian Kingsbury. 2013. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8599–8603.

Deng, Yonggang, and William Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 169–176.

——, and ——. 2008. HMM word and phrase alignment for statistical machine translation. *Audio, Speech, and Language Processing, IEEE Transactions on* 16.494–507.

Dhaka, Akash Kumar, and Giampiero Salvi. 2016. Semi-supervised learning with sparse autoencoders in phone classification. *ArXiv:1610.00520*.

Do, Thi-Ngoc-Diep, Alexis Michaud, and Eric Castelli. 2014a. Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a 'light' acoustic model of the target language and testing 'heavyweight' models from five national languages. In *4th International Workshop on Spoken Language Technologies for Under-resourced Languages*, 153–160.

Do, Van Hai, Xiong Xiao, Eng Siong Chng, and Haizhou Li. 2014b. Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages. *IEICE Transactions on Information and Systems* E97-D.285–295.

Duong, Long, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016a. An attentional model for speech translation without transcription. In *Proceedings of NAACL-HLT 2016*, 949–959.

——, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: cross-lingual parameter sharing in a neural network parser. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)* 845–850.

——, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016b. Learning crosslingual word embeddings without bilingual corpora. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* 1285–1295.

——, ——, ——, ——, and ——. 2017. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, 894–904.

Durrani, Nadir, Helmut Schmid, Alexander Fraser, and Philipp Koehn. 2014. Investigating the usefulness of generalized word representations in SMT. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, volume 14, 421–432.

Dyer, Chris. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 406–414.

——, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM Model 2. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologieslogies* 644–649.

El-Desoky Mousa, Amr, M. Ali Basha Shaik, Ralf Schluter, and Hermann Ney. 2010. Sub-lexical language models for German LVCSR. In *2010 IEEE Spoken Language Technology Workshop*, 171–176.

Elsner, Micha, Naomi H Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, number October, 42–54.

Fang, Meng, and Trevor Cohn. 2017. Model transfer for tagging low-resource languages using a bilingual dictionary. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 587–593.

Faruqui, Manaal, and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 462–471.

Feng, Yan-Mei, Li Xu, Ning Zhou, Guang Yang, and Shan-Kai Yin. 2012. Sine-wave speech recognition in a tonal language. *The Journal of the Acoustical Society of America* 131.133–138.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. In *ACM Transactions on Information Systems*, volume 20 of *WWW '01*, 116–131.

Frome, Andrea, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: a deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26*, 2121–2129.

Fung, Pascale, and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, 414–420.

Gandhe, A, F Metze, and I Lane. 2014. Neural network language models for low resource languages. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2615–2619.

Garofolo, John S, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue. 1993. TIMIT acoustic-phonetic continuous speech corpus. *Linguistic data consortium* 10.0.

Gauthier, Elodie, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in Sub-Saharan African languages for automatic speech recognition: A case study of Wolof. In *10th Language Resources and Evaluation Conference*, 3863–3867.

Ghahremani, Pegah, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. 2014. A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2494–2498.

Ghoshal, Arnab, Pawel Swietojanski, and Steve Renals. 2013. Multilingual training of deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 7319–7323.

Girshick, Ross. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

Godard, P., G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G-N. Kouarata, L. Lamel, H. Maynard, M. Mueller, A. Rialland, S. Stueker, F. Yvon, and M. Zanon-Boito. 2017. A very low resource language speech corpus for computational language documentation experiments. *ArXiv:1710.03501*.

Godard, Pierre, Gilles Adda, Martine Adda-decker, Alexandre Allauzen, Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kourata, Kevin Löser, Annie Rialland, and François Yvon. 2016. Preliminary experiments on unsupervised word discovery in Mboshi. In *Proceedings of the Annual Conference of the International Speech Communication Association, (INTERSPEECH)*.

Goldwater, Sharon, Thomas L Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 673–680.

——, Dan Jurafsky, and Christopher D Manning. 2008. Which words are hard to recognize? lexical, prosodic, and disfluency factors that increase ASR error rates. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL/HLT)*, 380–388.

Goodman, Joshua. 2001. A bit of progress in language modeling. *Technical Report* p. 73.

Gouws, Stephan, and Anders Sogaard. 2015. Simple task-specific bilingual word embeddings. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, 1302–1306.

Graff, David, Shudong Huang, Ingrid Cartagena, Kevin Walker, and Christopher Cieri, 2010. Fisher Spanish Transcripts LDC2010T04. Linguistic Data Consortium.

Graves, A, A.-R. Mohamed, and G Hinton. 2013. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6645–6649.

Graves, Alex, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd international conference on Machine Learning* 369–376.

——, and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. *JMLR Workshop and Conference Proceedings* 32.1764–1772.

Grézl, František, Martin Karafiát, and Miloš Janda. 2011. Study of probabilistic and bottle-neck features in multilingual environment. *Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding* 359–364.

Hadian, Hossein, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. 2018. Towards discriminatively trained HMM-based end-to-end models for automatic speech recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Haghighi, Aria, Percy Liang, T Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2008, 771–779.

Hanke, Florian R., 2017. *Computer supported collaborative language learning.* The University of Melbourne dissertation.

——, and Steven Bird. 2013. Large-scale text collection for unwritten languages. In *International Joint Conference on Natural Language Processing*, 1134–1138.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: scaling up end-to-end speech recognition. *ArXiv:1412.5567*.

Hao Fang, Mari Ostendorf, Peter Baumann, and Janet Pierrehumbert. 2015. Exponential language modeling using morphological features and multi-task learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.2410–2421.

Harris, Zellig S. 1954. Distributional structure. *Word* 10.146–162.

Harrison, K David. 2008. *When languages die: The extinction of the world's languages and the erosion of human knowledge*. Oxford University Press.

He, Yanzhang, and Eric Fosler-Lussier. 2012. Efficient segmental conditional random fields for phone recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association*, 1898–1901.

Heafield, Kenneth. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187–197.

Heck, M, Q T Do, S Sakti, G Neubig, T Toda, and S Nakamura. 2015. The NAIST ASR system for IWSLT 2015. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2015)*.

Heigold, Georg, Vincent Vanhoucke, Alan Senior, Patrick Nguyen, Marc'Aurelio Ranzato, Matthieu Devin, and Jeffrey Dean. 2013. Multilingual acoustic models using distributed deep neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 8619–8623.

Hermann, Karl Moritz, and Phil Blunsom. 2013. A simple model for learning multilingual compositional semantics. *ArXiv:1312.6173*.

Hermansky, Hynek. 1990. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America* 87.1738–1752.

Heyman, Geert, Ivan Vulić, and Marie-Francine Moens. 2017. Bilingual lexicon induction by learning to combine word-level and character-Level representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1085–1095.

Himmelmann, Nikolaus. 2006. Language documentation: what is it and what is it good for? In *Essentials of language documentation*, ed. by J Gippert, Nikolaus Himmelmann, and Ulrike Mosel, 1–30. Berlin/New York: de Gruyter.

Hinton, Geoffrey, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke,

Patrick Nguyen, Tara N Sainath, and Others. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE* 29.82–97.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9.1735–1780.

Hofmann, H, S Sakti, C Hori, H Kashioka, S Nakamura, and W Minker. 2012. Sequence-based pronunciation variation modeling for spontaneous ASR using a noisy channel approach. *IEICE Transactions on Information and Systems* 95.2084–2093.

Hu, W, Y Qian, and F K Soong. 2014. A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3206–3210.

Huang, Hank Chang Han, and Frank Seide. 2000. Pitch tracking and tone features for Mandarin speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 3.1523–1526.

Huang, Jui-Ting, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 7304–7308.

Hughes, Thad, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mike LeBeau. 2010. Building transcribed speech corpora quickly and cheaply for many languages. In *Interspeech 2010. Proceedings of the 11th Annual Conference of the International Speech Communication Association*, number September, 1914–1917.

Imseng, David, Petr Motlicek, Hervé Bourlard, and Philip N Garner. 2014. Using out-of-language data to improve an under-resourced speech recognizer. *Speech Communication* 56.142–151.

Irvine, Ann, and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 262–270.

Jansen, Aren, Kenneth Church, and Hynek Hermansky. 2010. Towards spoken term discovery at scale with zero resources. *Interspeech 2010* 1676–1679.

Jensson, Arnar, Koji Iwano, and Sadaoki Furui. 2008. Development of a speech recognition system for Icelandic using machine translated text. In *The first*

International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU-2008), 18–21.

Jensson, A.T. Arnar, Tasuku Oonishi, Koji Iwano, and Sadaoki Furui. 2009. Development of a WFST based speech recognition system for a resource deficient language using machine translation. *Proceedings of Asia-Pacific Signal and Information Processing Association* 50–56.

Jiang, Jie, Zeeshan Ahmed, Julie Carson-Berndsen, Peter Cahill, and Andy Way. 2011. Phonetic representation-based speech translation. In *13th Machine Translation Summit*.

Johnson, Mark. 2008. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, 20–27.

——, and Sharon Goldwater. 2009. Improving nonparametric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 317–325.

——, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. *Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics* 139–146.

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *ArXiv:1611.04558*.

Johnson, Samuel. 1755. Preface to the English Dictionary.

Jurafsky, Daniel, and James H Martin. 2009. *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Jyothi, Preethi, and Mark Hasegawa-Johnson. 2015. Acquiring speech transcriptions using mismatched crowdsourcing. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 1263–1269.

Kalchbrenner, Nal, and Phil Blunsom. 2013. Recurrent continuous translation models. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* 1700–1709.

Kamholz, David, Jonathan Pool, and Susan Colowick. 2014. PanLex: building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3145–3150.

Kamper, Herman, Aren Jansen, and Sharon Goldwater. 2017. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech and Language* 46.154–174.

Kempton, Timothy, and Roger K. Moore. 2014. Discovering the phoneme inventory of an unwritten language: a machine-assisted approach. *Speech Communication* 56.152–166.

Khadivi, Shahram, and Hermann Ney. 2008. Integration of speech recognition and machine translation in computer-assisted translation. *Audio, Speech, and Language Processing, IEEE Transactions on* 16.1551–1564.

Kim, Suyoun, Takaaki Hori, and Shinji Watanabe. 2016a. Joint CTC-attention based end-to-end speech recognition using multi-task learning. *ArXiv:1609.06773*.

Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016b. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2741–2749.

Kingma, Diederik P., and Jimmy Lei Ba. 2015. Adam: a method for stochastic optimization. *International Conference on Learning Representations 2015* 1–15.

Kirchhoff, Katrin, Dimitra Vergyri, Jeff Bilmes, Kevin Duh, and Andreas Stolcke. 2006. Morphology-based language modeling for conversational Arabic speech recognition. *Computer Speech and Language* 20.589–608.

Klementiev, Alexandre, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers*, number December, 1459–1474.

Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, 181–184.

Kočiský, Tomáš, Karl Moritz Hermann, and Phil Blunsom. 2014. Learning bilingual word representations by marginalizing alignments. *ArXiv:1405.0947*.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, volume 5, 79–86.

——. 2009. *Statistical machine translation.* Cambridge University Press.

——, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, and Others. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180.

——, and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition-Volume 9*, 9–16.

——, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, 48–54.

Köhler, Joachim. 1998. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 1, 417–420.

——. 1999. Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks. In *Workshop on Multi-Lingual Interoperability in Speech Technology (MIST).*

Krauss, Michael. 1992. The world's languages in crisis. *Language* 68.4–10.

Kunze, Julius, Louis Kirsch, Ilia Kurenkov, Andreas Krug, Jens Johannsmeier, and Sebastian Stober. 2017. Transfer learning for speech recognition on a budget. *ArXiv:1706.00290.*

Kurimo, Mikko, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi, and Tanel Alumäe. 2016. Modeling under-resourced languages for speech recognition. *Language Resources and Evaluation* 1–27.

Ladefoged, Peter. 1992. Another view of endangered languages. *Language* 68.809–811.

Laleye, Fréjus, Laurent Besacier, Eugène C. Ezin, and Cina Motamed. 2016. First automatic Fongbe continuous speech recognition system: development of acoustic models and language models. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*, volume 8, 477–482.

Lamel, L, J L Gauvain, V B Le, I Oparin, and S Meng. 2011. Improved models for Mandarin speech-to-text transcription. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4660–4663.

Landauer, Thomas K, and Susan T Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104.211.

Lankinen, Matti, Hannes Heikinheimo, Pyry Takala, Tapani Raiko, and Juha Karhunen. 2016. A character-word compositional neural language model for Finnish. *ArXiv:1612.03266*.

Lardilleux, Adrien, Julien Gosme, and Yves Lepage. 2010. Bilingual lexicon induction: effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, 252–256.

Lau, Jey Han, and Timothy Baldwin. 2016. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *1st Workshop on Representation Learning for NLP*, 78–86.

Le, Viet Bac, and Laurent Besacier. 2005. First steps in fast acoustic modeling for a new target language: application to Vietnamese. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 821–824.

Le, Viet-Bac, and Laurent Besacier. 2009. Automatic speech recognition for under-resourced languages: application to Vietnamese language. *Audio, Speech, and Language Processing, IEEE Transactions on* 17.1471–1482.

Lee, Chia-ying, and James Glass. 2012. A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL)*, 40–49.

——, Timothy J O'Donnell, and James Glass. 2015. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics* 3.389–403.

Lee, Jason, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *ArXiv:1610.03017*.

Lee, Kai Fu. 1990. Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38.599–609.

Lee, Tan, Wai Lau, Yiu Wing Wong, and P C Ching. 2002. Using tone information in Cantonese continuous speech recognition. *ACM Transactions on Asian Language Information Processing (TALIP)* 1.83–102.

Lee, Young-Suk. 2004. Morphological analysis for statistical machine translation. In *HLT-NAACL 2004: Short Papers*, 57–60.

Lei, Xin, Manhung Siu, Mei-Yuh Hwang, Mari Ostendorf, and Tan Lee. 2006. Improved tone modeling for Mandarin broadcast news speech recognition. *Proceedings of the International Conference on Spoken Language Processing* 1237–1240.

Levin, Keith, Katharine Henry, Aren Jansen, and Karen Livescu. 2013. Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 410–415.

Levow, Gina-Anne, Douglas W Oard, and Philip Resnik. 2005. Dictionary-based techniques for cross-language information retrieval. *Information processing & management* 41.523–547.

Lewis, M Paul, and Gary F Simons. 2010. Assessing endangerement: expanding Fishman's GIDS. *Revue roumaine de linguistique* 55.2.

——, ——, and Charles D Fennig (eds.). 2015. *Ethnologue: Languages of the World, Eighteenth edition.* Dallas, Texas: SIL International.

Li, Xiangang, and Xihong Wu. 2016. Long short-term memory based convolutional recurrent neural networks for large vocabulary speech recognition. *ArXiv:1610.03165*.

Li, Zezhong, Hideto Ikeda, and Junichi Fukumoto. 2013. Bayesian word alignment and phrase table training for statistical machine translation. *IEICE TRANSACTIONS on Information and Systems* 96.1536–1543.

Lidz, Liberty A, 2010. *A descriptive grammar of Yongning Na (Mosuo).* University of Texas, Department of linguistics dissertation.

Lin, Hui, Li Deng, Dong Yu, Yi-fan Gong, Alex Acero, and Chin-Hui Lee. 2009. A study on multilingual acoustic modeling for large vocabulary ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2009*, 4333–4336.

Ling, Wang, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *ArXiv:1511.04586*.

Liu, Chunxi, Preethi Jyothi, Hao Tang, Vimal Manohar, Rose Sloan, Tyler Kekona, Mark Hasegawa-johnson, and Sanjeev Khudanpur. 2016. Adapting ASR for under-resourced languages using mismatched transcriptions. *ICASSP 2016* 5840–5844.

——, Jinyi Yang, Ming Sun, Santosh Kesiraju, Alena Rott, Lucas Ondel, Pegah Ghahremani, Najim Dehak, Lukas Burget, and Sanjeev Khudanpur. 2017a. An empirical evaluation of zero resource acoustic unit discovery. *ArXiv:1702.01360*.

Liu, Hairong, Zhenyao Zhu, Xiangang Li, and Sanjeev Satheesh. 2017b. Gram-CTC: automatic unit selection and target decomposition for sequence labelling. *ArXiv:1703.00096*.

Lu, Liang, Lingpeng Kong, Chris Dyer, and Noah A. Smith. 2017. Multitask learning with CTC and segmental CRF for speech recognition. *ArXiv:1702.06378*.

——, ——, ——, ——, and Steve Renals. 2016. Segmental recurrent neural networks for end-to-end speech recognition. *ArXiv:1603.00223*.

Luong, Minh-Thang, and Min-Yen Kan. 2010. Enhancing morphological alignment for translating highly inflected languages. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 743–751.

Ma, Guojie, Xingshan Li, and Keith Rayner. 2014. Word segmentation of overlapping ambiguous strings during Chinese reading. *12* 40.1046.

Maas, Andrew L, Ziang Xie, Dan Jurafsky, and Andrew Y Ng. 2015. Q. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mariño, José B, Rafael E Banchs, Josep M Crego, Adrià De Gispert, Patrick Lambert, José A R Fonollosa, and Marta R Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics* 32.527–549.

Martens, James. 2011. Generating text with recurrent neural networks. In *Neural Networks*, volume 131, 1017–1024.

Matusov, Evgeny, Stephan Kanthak, and Hermann Ney. 2005. On the integration of speech recognition and statistical machine translation. In *INTERSPEECH*, 3177–3180.

Melamed, I Dan. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, 125–134.

Mermer, Coşkun, and Murat Saraçlar. 2011. Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 182–187.

——, Murat Saraçlar, and Ruhi Sarikaya. 2013. Improving statistical machine translation using Bayesian word alignment and Gibbs sampling. *Audio, Speech, and Language Processing, IEEE Transactions on* 21.1090–1101.

Metze, Florian, Zaid A.W. W Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, and Van Huy Nguyen. 2013. Models of tone for tonal and non-tonal languages. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2013 - Proceedings* 261–266.

Miao, Yajie, Mohammad Gowayyed, and Florian Metze. 2015. EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding. *ArXiv:1507.08240*.

——, and Florian Metze. 2017. End-to-end architectures for speech recognition. In *New Era for Robust Speech Recognition: Exploiting Deep Learning*, ed. by Shinji Watanabe, Marc Delcroix, Florian Metze, and John R Hershey, 299–323. Cham: Springer International Publishing.

Michailovsky, Boyd, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation* 8.119–135.

Michaud, Alexis. 2008. Phonemic and tonal analysis of Yongning Na*. *Cahiers de Linguistique Asie Orientale* 37.159–196.

——, 2016. Online Na-English-Chinese Dictionary. *Halshs-01204638*. https://halshs.archives-ouvertes.fr/halshs-01204638.

——, 2017a. Speech recognition for newly documented languages: highly encouraging tests using automatically generated phonemic transcription of Yongning Na audio recordings. *Himalco*. https://himalco.hypotheses.org/285.

——. 2017b. *Tone in Yongning Na: lexical tones and morphotonology*. Number 13 in Studies in Diversity Linguistics. Berlin: Language Science Press.

——, and Dashilame Latami, 2017. Housebuilding 2. *Pangloss Collection.* http://lacito.vjf.cnrs.fr/pangloss/corpus/ show_text_en.php?id=crdo-NRU_F4_HOUSEBUIL DING2_SOUND&idref=crdo-NRU_F4_HOUSEBU ILDING2.

Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representation in vector space. In *ICLR*.

——, M Karafiat, L Burget, J Cernocky, and S Khudanpur. 2010. Recurrent neural network based language model. In *Proceedings of the Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 1045–1048.

——, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. *ArXiv:1309.4168*.

——, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

——, Ilya Sutskever, Anoop Deoras, Hai-Son Le, Stefan Kombrink, and Jaň Cernock. 2012. Subword language modeling with neural networks.

Miranda, Joao, Joao P. Neto, and Alan W. Black. 2012a. Parallel combination of speech streams for improved ASR. In *Interspeech-2012*, 2–5.

——, Joao Paulo Neto, and Alan W Black. 2012b. Recovery of acronyms, out-of-lattice words and pronunciations from parallel multilingual speech. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 348–353.

Miyaoka, Osahito, Osamu Sakiyama, and Michael E Krauss. 2007. *The vanishing languages of the Pacific Rim.* Oxford University Press, USA.

Mnih, Andriy, Zhang Yuecheng, and Geoffrey Hinton. 2009. Improving a statistical language model through non-linear prediction. *Neurocomputing* 72.1414–1418.

Mochihashi, Daichi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 100–108.

Mortensen, David R, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: a resource for mapping IPA

segments to articulatory feature vectors. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* 3475–3484.

Müller, Markus, Sebastian Stüker, and Alex Waibel. 2017. Language Adaptive Multilingual CTC Speech Recognition. In *Speech and Computer: 19th International Conference, SPECOM 2017, Hatfield, UK, September 12-16, 2017, Proceedings*, 473–482.

Nagata, Masaaki. 1997. A self-organizing Japanese word segmenter using heuristic word identification and re-estimation. In *Proceedings of the 5th Workshop on Very Large Corpora*, 203–215.

Nakov, Preslav, and Hwee Tou Ng. 2011. Translating from morphologically complex languages: a paraphrase-based approach. In *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, volume 80, 1298–1307.

——, and Jörg Tiedemann. 2012. Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, 301–305.

Nallasamy, Udhyakumar, Florian Metze, and Tanja Schultz. 2012. Semi-supervised learning for speech recognition in the context of accent adaptation. In *Proceedings of Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, 2–6.

Neubig, Graham, 2012. *Unsupervised learning of lexical information for language processing systems*. Kyoto University dissertation.

——. 2013. Travatar: a forest-to-string machine translation engine based on tree transducers. In *ACL (Conference System Demonstrations)*, 91–96.

——. 2014. Simple, correct parallelization for blocked Gibbs sampling. Technical report, Nara Institute of Science and Technology.

——, and Chris Dyer. 2016. Generalizing and hybridizing count-based and neural language models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1163–1172.

——, Masato Mimura, and Tatsuya Kawahara. 2012a. Bayesian learning of a language model from continuous speech. *IEICE TRANSACTIONS on Information and Systems* 95.614–625.

——, Yosuke Nakata, and Shinsuke Mori. 2011a. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 529–533.

——, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012b. Machine translation without words through substring alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 165–174.

——, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. 2011b. An unsupervised model for joint phrase alignment and extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 632–641.

Ney, Hermann. 1999. Speech translation: Coupling of recognition and translation. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 1, 517–520.

Ng, Raymond W M, Thomas Hain, and Trevor Cohn. 2013. Adaptation of lecture speech recognition system with machine translation output. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 8401–8405.

Nguyen, Thuylinh, Stephan Vogel, and Noah A Smith. 2010. Nonparametric word segmentation for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 815–823.

Nguyen, Toan Q., and David Chiang. 2017. Transfer learning across low-resource, related Languages for neural machine translation. *ArXiv:1708.09803*.

O. Abdel-Hamid L. Deng, D Yu, and H Jiang. 2013. Deep segmental neural networks for automatic speech recognition. In *Proceedings of Interspeech*, p. 70.

Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29.19–51.

Osband, Ian, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep exploration via bootstrapped DQN. *ArXiv:1602.04621*.

Östling, Robert, and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *ArXiv:1708.05729*.

Park, Alex S, and James R Glass. 2008. Unsupervised pattern discovery in speech. *Audio, Speech, and Language Processing, IEEE Transactions on* 16.186–197.

Paulik, Matthias, Sebastian Stüker, C Fugen, Tanja Schultz, Thomas Schaaf, and Alex Waibel. 2005. Speech translation enhanced automatic speech recognition. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, 121–126.

——, and Alex Waibel. 2009. Automatic translation from parallel speech: Simultaneous interpretation as MT training data. In *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2009*, 496–501.

——, and ——. 2010. Spoken language translation from parallel speech audio: simultaneous interpretation as SLT training data. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 5210–5213.

——, and ——. 2013. Training speech translation from audio recordings of interpreter-mediated communication. *Computer Speech & Language* 27.455–474.

Pelemans, Joris, Tom Vanallemeersch, Kris Demuynck, Patrick Wambacq, and Others. 2015. Efficient language model adaptation for automatic speech recognition of spoken translations. In *Proceedings of the Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 2262–2266.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. Glove: global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12.1532–1543.

Plahl, Christian, Ralf Schlüter, and Hermann Ney. 2011. Cross-lingual portability of Chinese and English neural network features for French and German LVCSR. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, 371–376.

Post, Matt, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the Fisher and Callhome Spanish–English speech translation corpus. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77.257–286.

Ramsundar, Bharath, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. 2015. Massively multitask networks for drug discovery. *ArXiv:1502.02072*.

Reddy, Aarthi, and Richard C Rose. 2010. Integration of statistical models for dictation of document translations in a machine-aided human translation task. *Audio, Speech, and Language Processing, IEEE Transactions on* 18.2015–2027.

Redeker, Gisela. 1984. On differences between spoken and written language. *Discourse Processes* 7.43–55.

Reiman, Will D. 2010. Basic oral language documentation. *Language Documentation & Conservation* 254–268.

Rodriguez, Luis, Aarthi Reddy, and Richard Rose. 2012. Efficient integration of translation and speech models in dictation based machine aided human translation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 4949–4952.

Ruder, Sebastian. 2017. An overview of multi-task learning in deep neural networks. *ArXiv:1706.05098*.

Sak, Hasim, Murat Saraclar, and Tunga Gungor. 2010. Morphology-based and sub-word language modeling for Turkish speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5402–5405.

Sak, Haşim, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Sakti, Sakriani, Andrew Finch, Chiori Hori, Hideki Kashioka, and Satoshi Nakamura. 2011. Conditional random fields for modeling Korean pronunciation variation. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, 49–55.

Sam, Sethserey, Eric Castelli, and Laurent Besacier. 2012. Online unsupervised multilingual acoustic model adaptation for nonnative ASR. *ASEAN Engineering Journal* 1.76–86.

Sassano, Manabu. 2014. Deterministic word segmentation using maximum matching with fully lexicalized rules. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* 79–83.

Scharenborg, Odette, Francesco Ciannella, Shruti Palaskar, Alan Black, Florian Metze, Lucas Ondel, and Mark Hasegawa-Johnson. 2017. Building an ASR system for a low-research language through the adaptation of a high-resource language ASR system: preliminary results. In *International Conference on Natural Language, Signal and Speech Processing*.

Schröder, Marc, and Jürgen Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology* 6.365–377.

Schultz, Tanja, and Alex Waibel. 2001a. Experiments on cross-language acoustic Modeling. *Proc. EUROSPEECH'01* 2721–2724.

——, and ——. 2001b. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication* 35.31–51.

Schuster, Mike, and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45.2673–2681.

Scott, Steven L. 2002. Bayesian methods for hidden Markov models. *Journal of the American Statistical Association* 97.337–351.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *ArXiv:1508.07909*.

Sercu, Tom, Christian Puhrsch, Brian Kingsbury, and Yorktown Heights. 2016. Very deep multilingual convolutional neural networks for LVCSR. In *ICASSP 2016*, 4955–4959.

Shaik, M Ali Basha, El-Desoky Mousa, Ralf Schlüter, and Hermann Ney. 2011. Hybrid language models using mixed types of sub-lexical units for open vocabulary German LVCSR.

Shan, Changhao, Junbo Zhang, Yujun Wang, and Lei Xie. 2017. Attention-based end-to-end speech recognition in Mandarin. *ArXiv:1707.07167*.

Shareghi, Ehsan, Matthias Petri, Gholamreza Haffari, and Trevor Cohn. 2016. Fast, small and exact: infinite-order language modelling with compressed suffix trees. *ArXiv:1608.04465*.

Shazeer, Noam, Ryan Doherty, Colin Evans, and Chris Waterson. 2016. Swivel: improving embeddings by noticing what's missing. *ArXiv:1602.02215*.

Simons, Gary F, and Charles D Fennig (eds.) 2017. *Ethnologue: languages of the world*. Dallas: SIL International, twentieth edition.

Soltau, Hagen, Hank Liao, and Hasim Sak. 2016. Neural speech recognizer: acoustic-to-word LSTM model for large vocabulary speech recognition. *ArXiv:1610.09975*.

Sriram, Anuroop, Heewoo Jun, Yashesh Gaur, and Sanjeev Satheesh. 2017. Robust speech recognition using generative adversarial networks. *ArXiv:1711.01567*.

Stahlberg, F, T Schlippe, S Vogel, and T Schultz. 2015. Cross-lingual lexical language discovery from audio data using multiple translations. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 5823–5827.

Stahlberg, Felix, Tim Schlippe, Stephan Vogel, and Tanja Schultz. 2013. Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. In *Statistical Language and Speech Processing*, 260–272. Springer.

——, ——, ——, and ——. 2014a. Towards automatic speech recognition without pronunciation dictionary, transcribed speech and text resources in the target language using cross-lingual word-to-phoneme alignment. In *The 4th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU)*, 73–80.

——, ——, ——, and ——. 2014b. Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. *Computer Speech & Language* 234–261.

——, Tim Schlippe, Sue Vogel, and Tanja Schultz. 2012. Word segmentation through cross-lingual word-to-phoneme alignment. In *Spoken Language Technology Workshop (SLT), 2012*, 85–90.

Stolcke, A., F. Grezl, Mei-Yuh Hwang, Xin Lei, N. Morgan, and D. Vergyri. 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 1, I–321–I–324.

Strassel, Stephanie, and Jennifer Tracey. 2016. LORELEI language packs: data, tools, and resources for technology development in low resource languages. *LREC* 10–11.

Stüker, Sebastian, Laurent Besacier, and Alex Waibel. 2009. Human translations guided language discovery for ASR systems. In *INTERSPEECH*, 3023–3026.

——, Kevin Kilgour, and Florian Kraft. 2012. Quaero 2010 speech-to-text evaluation systems. In *High Performance Computing in Science and Engineering '11*, 607–618.

——, and Alex Waibel. 2008. Towards human translations guided language discovery for ASR systems. In *SLTU*, 76–79.

Su, Jinsong, Zhixing Tan, Deyi Xiong, Rongrong Ji, Xiaodong Shi, and Yang Liu. 2016. Lattice-based recurrent neural network encoders for neural machine translation. *ArXiv:1609.07730*.

Sun, Weiwei. 2011. A stacked sub-word model for joint Chinese word segmentation and part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1385–1394.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, 3104–3112.

Takezawa, Toshiyuki, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *LREC*, 147–152.

Tang, Hao, Liang Lu, Lingpeng Kong, Kevin Gimpel, Karen Livescu, Chris Dyer, Noah A Smith, and Steve Renals. 2017. End-to-end neural segmental models for speech recognition. *IEEE Journal of Selected Topics in Signal Processing* 11.8.1254–1264.

Teh, Yee Whye. 2006. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 985–992.

Thieberger, Nick. 2016. Documentary linguistics: methodological challenges and innovatory responses. *Applied Linguistics* 37.88–99.

Thomas, Samuel, Sriram Ganapathy, Hynek Hermansky, and Speech Processing. 2012. Multilingual MLP features for low-resource LVCSR systems. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 4269–4272.

——, Michael L Seltzer, Kenneth Church, and Hynek Hermansky. 2017. Deep neural network features and semi-supervised training for low-resource speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*.

Tiedemann, Jörg. 2009. Character-based PSMT for closely related languages. *Proceedings of 13th Annual Conference of the European Association for Machine Translation* 9.12–19.

Tietz, Marian, Tayfun Alpay, Johannes Twiefel, and Stefan Wermter. 2017. Semi-supervised phoneme recognition with recurrent ladder networks. *ArXiv:1706.02124*.

Toshniwal, Shubham, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2017. Multilingual speech recognition with a single end-to-end model. *ArXiv:1711.01694*.

Tóth, László, Joe Frankel, Gábor Gosztolya, and Simon King. 2008. Cross-lingual portability of MLP-based tandem features - A case study for english and Hungarian. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2695–2698.

Varadarajan, Balakrishnan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of ACL-08: HLT, Short Papers*, 165–168.

Verwimp, Lyan, Joris Pelemans, Hugo Van hamme, and Patrick Wambacq. 2017. Character-word LSTM language models. *ArXiv:1704.02813*.

Vidal, Enrique, Francisco Casacuberta, Luis Rodriguez, Jorge Civera, and Carlos D Martínez Hinarejos. 2006. Computer-assisted translation using speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on* 14.941–951.

Vidal, Exirique. 1997. Finite-state speech-to-speech translation. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 1, 111–114.

Vilar, David, Jan-T Peter, and Hermann Ney. 2007. Can we translate letters? In *Proceedings of the Second Workshop on Statistical Machine Translation*, 33–39.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, 836–841.

Vu, Ngoc Thang, David Imseng, Daniel Povey, Petr Motlicek, Tanja Schultz, and Hervé Bourlard. 2014. Multilingual deep neural network based acoustic modeling for rapid language adaptation. In *Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7639–7643.

——, Florian Metze, and Tanja Schultz. 2012. Multilingual bottle-neck features and its application for under-resourced languages. In *The third International Workshop on Spoken Language Technologies for Under-resourced languages*.

Vulic, Ivan, and Marie-Francine Moens. 2015. Bilingual word embeddings from non-parallel document-aligned data applied to bilingual lexicon induction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 719–725.

Vulić, Ivan, Roy Schwartz, Ari Rappoport, Roi Reichart, and Anna Korhonen. 2016. Automatic selection of context configurations for improved (and fast) class-specific word representations. *ArXiv:1608.05528*.

Wang, Rui, Hai Zhao, Sabine Ploux, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2016. A novel bilingual word embedding method for lexical translation using bilingual sense clique. *ArXiv:1607.08692*.

Wang, Yining, Long Zhou, Jiajun Zhang, and Chengqing Zong. 2017a. Word, subword or character? An empirical study of granularity in Chinese-English NMT. *ArXiv:1711.04457*.

Wang, Yisen, Xuejiao Deng, Songbai Pu, and Zhiheng Huang. 2017b. Residual convolutional CTC networks for automatic speech recognition. *ArXiv:1702.07793*.

Weiss, Ron J., Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly transcribe foreign speech. *ArXiv:1703.08581*.

Wheatley, Barbara, 1996. CALLHOME Spanish Transcripts LDC96T17. Linguistic Data Consortium.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics* 23.377–403.

——, and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 206–213.

Xiao, Min, and Yuhong Guo. 2014. Distributed word representation learning for cross-lingual dependency parsing. In *CoNLL*, 119–129.

Xu, Haihua, Van Hai Do, Xiong Xiao, and Eng Siong Chng. 2015. A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2132–2136.

——, Hang Su, Chongjia Ni, Xiong Xiao, Hao Huang, Eng-Siong Chng, and Haizhou Li. 2016. Semi-supervised and cross-lingual knowledge transfer learnings for DNN hybrid acoustic models under low-resource conditions. In *Proceedings of the Annual Conference of the International Speech Communication Association, (INTERSPEECH)*, 1315–1319.

Xu, Jia, Jianfeng Gao, Kristina Toutanova, and Hermann Ney. 2008. Bayesian semi-supervised Chinese word segmentation for statistical machine translation. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, 1017–1024.

——, Richard Zens, and Hermann Ney. 2004. Do we need Chinese word segmentation for statistical machine translation. In *Proceedings of the Third SIGHAN Workshop on Chinese Language Processing*, 257–264.

Xu, Ping, and Pascale Fung. 2013. Cross-lingual language modeling for low-resource speech recognition. *IEEE Transactions on Audio, Speech and Language Processing* 21.1134–1144.

Yamada, Kenji, and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, 523–530.

Yang, Zhenhong, Liberty A Lidz, and Others. 2009. An overview of the Mosuo language. *Linguistics of the Tibeto-Burman Area* 32.1.

Yu, Seunghak, Nilesh Kulkarni, Haejun Lee, and Jihie Kim. 2017. Syllable-level neural language model for agglutinative language. *ArXiv:1708.05515*.

Zaremba, Wojciech, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *ArXiv:1409.2329*.

Zenkel, Thomas, Ramon Sanabria, Florian Metze, Jan Niehues, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2017. Comparison of decoding strategies for CTC acoustic models. *ArXiv:1708.04469*.

Zhang, Hao, Chris Quirk, Robert C Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *ACL*, 97–105.

Zhang, Jiajun, Shujie Liu, Mu Li, Ming Zhou, and Chengqing Zong. 2014. Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 111–121.

Zhang, Ying, Mohammad Pezeshki, Yoshua Bengio, Aaron Courville, and Q C Hc. 2016a. Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks. *ArXiv:1701.02720*.

——, Mohammad Pezeshki, Philémon Brakel, Saizheng Zhang, César Laurent, Yoshua Bengio, and Aaron Courville. 2016b. Towards end-to-end speech recognition with deep convolutional neural networks. *ArXiv:1701.02720*.

Zhang, Zewang, Zheng Sun, Jiaqi Liu, Jingwen Chen, Zhao Huo, and Xiao Zhang. 2016c. Deep recurrent convolutional neural network: improving performance for speech recognition. *ArXiv:1611.07174*.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *ArXiv:1604.02201*.

Zou, Will Y, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, 1393–1398.

Zwarts, Simon, and Mark Dras. 2007. Statistical machine translation of australian aboriginal languages: morphological analysis with languages of differing morphological richness. In *Proceedings of the Australasian Language Technology Workshop*, 134–142.

Zweig, G., C. Yu, J. Droppo, and A. Stolcke. 2016. Advances in all-neural speech recognition. *ArXiv:1609.05935*.

Řehůřek, Radim, and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.