Interleaved Speech for Language Documentation

Name of author

Address - Line 1
Address - Line 2
Address - Line 3

Abstract

The world's languages are dying out, and there is little linguistic record of most of them. Both the collection of data and the efficient processing of such data are key bottlenecks hindering the creation of a record of all human languages. This thesis is an investigation of natural language processing of bilingual speech in order to better cope with such languages. In particular, source language speech *interleaved* with spoken target translations is considered. The tasks of lexicon induction and phonemic transcription of unwritten languages are investigated as they are key steps in language documentation. These tasks are difficult since data is limited, with quality acoustic models and language models being unavailable. This motivates an investigation into approaches that harness the translation in a large target language to disambiguate the source signal. In addition to tackling these tasks, the PhD project aims to investigate how such models may help further guide data acquisition in order to make the language documentation process more efficient.

Index Terms: speech recognition, machine translation, low-resource languages

1. Introduction

1.1. Language Death

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 [1]. Most of the approximately 7,000 languages catalogued by the Ethnologue project [2] have no orthography and thus no written record. Unless these languages are documented before they die out, much linguistic, cultural and anthropological information will be lost forever.

1.2. Language Documentation 2.0

This threat of language death motivates field linguists to engage in the documentation of languages. Traditionally this involves the linguist traveling to remote communities for one-on-one elicitation of data from speakers of threatened languages in order to prepare text collections, lexicons and grammars of their languages. However, this process is slow and there are a limited number of linguists engaged in it. Given the estimated rate of language death, it's clear that the current rate of collection is insufficient in order to adequately document most of the world's languages before they die out.

The proliferation of cheap mobile phones is creating new opportunities for documentating languages in a manner more efficient than traditional approaches [3, 4, 5]. *Aikuma* is one such app that aims to provide field linguists with greater leverage in eliciting speech data through the use of a crowdsourcing model. Since most speakers of endangered languages are bilingual, Aikuma aims to elicit bilingual *interleaved speech*, which consists of segments (usually at the phrase or sentence level) of endangered source speech paired with spoken translations in larger language, the latter of the two being a language that can be more reliably transcribed.

This data has specific features that distinguish it from the data used in most natural language processing research. It consists of limited quantities of bilingual spoken data. The segments are of varying size, ranging between words and paragraphs as determined by the translator.

1.3. Research Questions

There are two broad research questions this setting gives rise to:

1. How can we use this data to improve the performance of fundamental tasks such as phoneme transcription and lexicon induction?
2. How can this information further guide the language documentation process so that the limited resources can be better utilized?

This thesis focuses primarily on the first question, by tackling the tasks of bilingual lexicon induction (Section 2) and improved phoneme recognition (Section 3). These are addressed first in artificial settings, before adaptation to real-world scenarios (Section 4.1). However the two broad questions are interrelated and one aspect of the investigation bridges the gap by investigating how the models may better inform the data collection process (Section 4.2).

2. Bilingual Lexicon Induction

The first task is that of determining bilingual word pairs given a sequence of source language phonemes. Preliminary experiments demonstrate that a competitive end-to-end machine translation system can be built using parallel data consisting of source phonemes and target words when sufficient data is available. Though translation performance is completely inadequate in the face of little data, the most confident entries in the phrase tables are frequently correct, even when such limited data is available. In [6], a collection of bilingual lexicon induction techniques are investigated to assess their performance in the face of limited data, demonstrating bilingual lexical entries can be determined with high precision with as little as 1,000 sentences of parallel text. Figure 1 compares the output lexicons of three methods against a GIZA++ [7] baseline. *Bayes ITG* acquires lexical entries using pialign [8]; *UWS GIZA++* performs unsupervised word segmentation before alignment with GIZA++; while *Model 3P* uses the model of [9]. The precision

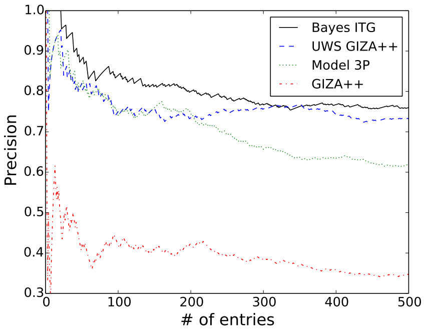

Figure 1: Comparison of methods of bilingual lexicon induction precisions over a 10k sentence dataset.

at K is presented, using strict criteria on the accuracy of a bilingual lexical item. Many that were considered not strictly correct were nonetheless meaningful phrase pairs.

3. Improving Phoneme Recognition

The previous work assumed accurate phonemic transcription, which is unrealistic given the circumstances. Though some of the techniques have previously proven resilient to some degree of noise [10], the problem of obtaining accurate phonemic transcription affects the models significantly, and is an important step of language documentation in its own right.

3.1. Exploring phoneme classes

A preliminary exploration into the use of equivalence classes to represent how phonemes are confusable with one another was undertaken. The hypothesis was that by grouping phonemes into such classes, a voting algorithm could then be used to determine what phonemes are likely in different contexts. In practice, however, phonemes simply do not fit into neat classes that capture their confusability. Insertions and deletions also frequently occur and could not be captured by the model.

3.2. Learning directly from lattices

Rather than using a one-best ASR hypothesis, another method was undertaken that involves learning directly from lattices. As a first step, and a useful investigation in its own right, learning from word lattices was investigated instead of phoneme lattices, though this presupposes the availability of a lexicon. By composing the lattice with a weighted finite-state transducer representing a lexical translation model, Bayesian inference can be performed to determine the translation model parameters, in turn allowing for better speech recognition, as presented in [11]. Figure 2 shows how such an approach can improve word error rates even when training data is very limited. Further work demonstrates this principle can be applied to phoneme lattices in a non-parameteric Bayesian framework that additionally segments and learns a lexicon.

4. Real world low-resource languages

The above investigations used artificial data from well-resourced languages. An important goal of this PhD project is to apply these methods to real-world scenarios.

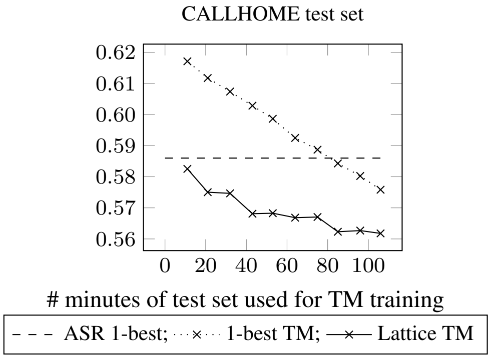

Figure 2: Word error rates of the method, Lattice TM versus two baselines.

4.1. Application of the above techniques to real languages

The primary problem faced in applying the above approaches to real low-resource languages is the quality of the suitability of the acoustic model used to generate the phoneme lattices. The most promising approaches are to take multilingual acoustic models (universal recognizers) or acoustic models of similar languages, with performance largely dependent on the specific language in question. Additional challenges such as heavy tonality of some languages may amplify the problem [12].

On the flip side, there is often some amount of prior information that can be used to inform the model. In addition to small quantities of phonemically transcribed data for acoustic model adaption, linguists typically have knowledge of the phoneme inventory, and some sort of limited lexicon and possibly further grammatical information. Seeding the model described in Section 3.2 with lexical items we are sure about may help overcome some of the difficulties. Extending the models to make the most of all the prior information available will be essential in a real-world scenario.

4.2. Speaker confirmation

With the expectation that the discussed methods will not be able to solve all problems in light of insufficient data, perhaps one of the most promising applications of the methods will be in guiding the efficient data collection, since person-hours are very limited.

One promising avenue is to seek confirmation from mother-tongue speakers of pieces of information inferred by the model. The speakers may be presented with snippets of source audio corresponding to hypothesized source words, and a written list of likely target alignments that they can either confirm or refute. Confirming the segmentation and alignment of the audio can both improve the transcription directly through the harnessing of the translation model as well as improving the translation model itself. The approach may be extended to additionally collect respeakings of segments the model is less sure about. Thus, the manual confirmation and automatic inference may be paired in an iterative framework where the inference informs what aspects of the data require further attestations, and the confirmations aid further inference.

5. References

[1] O. Miyaoka, O. Sakiyama, and M. E. Krauss, *The vanishing languages of the Pacific Rim*. Oxford University Press, USA, 2007.

[2] M. P. Lewis, G. F. Simons, and C. D. F. (eds.), *Ethnologue: Languages of the World, Eighteenth edition*. Dallas, Texas: SIL International, 2015. [Online]. Available: Online version: http://www.ethnologue.com

[3] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. J. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages." in *INTERSPEECH*, 2010, pp. 1914–1917.

[4] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014.

[5] S. Bird, F. R. Hanke, O. Adams, and H. Lee, "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*. Baltimore, Maryland, USA: ACL, jun 2014, pp. 1–5.

[6] O. Adams, G. Neubig, T. Cohn, and S. Bird, "Inducing Bilingual Lexicons from Small Quantities of Sentence-Aligned Phonemic Transcriptions," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, dec 2015.

[7] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[8] G. Neubig, T. Watanabe, S. Mori, and T. Kawahara, "Machine Translation without Words through Substring Alignment," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Jeju Island, Korea: ACL, jul 2012, pp. 165–174.

[9] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, "Word segmentation through cross-lingual word-to-phoneme alignment," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 85–90.

[10] ——, "Pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment," in *Statistical Language and Speech Processing*. Springer, 2013, pp. 260–272.

[11] O. Adams, G. Neubig, T. Cohn, and S. Bird, "Learning a translation model from word lattices," in *17th Annual Conference of the International Speech Communication Association (InterSpeech 2016)*, San Francisco, California, USA, September 2016.

[12] T.-N.-D. Do, A. Michaud, and E. Castelli, "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 (SLTU 2014)*, St Petersburg, Russia, may 2014, pp. 153–160. [Online]. Available: https://halshs.archives-ouvertes.fr/halshs-00980431