

TOWARDS SPEECH TRANSLATION OF NON WRITTEN LANGUAGES

Laurent Besacier, Bowen Zhou, Yuqing Gao*

IBM T.J. Watson Research Center

1101 Old Kitchawan Road, Yorktown Heights, NY, 10598, USA

* also at CLIPS-IMAG Laboratory, CNRS, BP 53, 38041 Grenoble Cedex 9, FRANCE

ABSTRACT

A large amount of languages in the world do not have an acknowledged written form. However, for a task like speech to speech translation, the written form of a language may be considered as secondary and it might be possible, under certain conditions, to bypass it. This paper is our first attempt to show that such an approach is possible. We propose a phone-based speech translation approach where translation models are learned on a parallel corpus made of foreign phone sequences and their corresponding English translation. Our experiments show that using our so-called phone-based approach leads almost to the same performance as the baseline approach, while being theoretically applicable to any non written language.

Index Terms— speech to speech translation, non written languages, unsupervised word discovery

1. INTRODUCTION

According to [1], a large number of languages in the world do not have an acknowledged written form and only 5-10% of all languages use one of about 25 writing systems¹. This observation is also confirmed in [2] which claims that among the languages worldwide, most of them are not written nor even described in the academic literature. These include first many indigenous languages without a literary tradition but also many dialects used for everyday conversations but not for written communication. Chinese and Arabic are examples of languages with a large number of unwritten regional dialects that differ significantly from the standard language [1]. In that case, the collection of resources for spoken language technology is more difficult since the only way to obtain “text” material is to record and transcribe data in that language. This may require the definition of a new writing system and orthographic standards as done in recent DARPA projects (EARS for Levantine Arabic, and TRANSTAC for Iraqi Arabic). While this approach may be usable for dialects related to a standard written language, it might be far more painful or even impossible to apply to other languages that are not related to

any written standard. Should we just give up in developing SLT systems for these languages? We rather think that for a task like speech to speech translation, the written form of a language may be considered as secondary and that it might be possible, under certain conditions, to bypass the written form to perform speech translation. This paper is our first attempt to show that such an approach is possible. We try to build a foreign to english speech translation system with parallel data consisting of foreign audio recordings of a non written language and their corresponding English text translation. In addition, we make the assumption that the foreign training speech signals are manually transcribed in phonetic sequences. While such a transcription (using for instance the *International Phonetic Alphabet*) may be more time-consuming than a standard word transcription², we believe that it is still realistic. The task of the human annotator may be for instance reduced by first applying a multilingual phone recognizer on the foreign speech data.

In short, this paper proposes to compare two different approaches for speech translation of non written languages:

-1- a *baseline* approach where a writing system is defined for the foreign language (as done in DARPA TRANSTAC project), and the speech data is transcribed (into words) and translated to obtain parallel english / foreign text,

-2- a *phone-based* approach where the speech data is transcribed into phonemes sequences, and translated; the parallel corpus is then made of foreign phone sequences and their corresponding English text.

In both cases, we use phrase-based SMT which is currently one of the best performing methods in statistical machine translation and might be particularly suitable here : while in case -1-, the foreign input is segmented into arbitrary multi-word units (so-called “phrases”); in case -2- the foreign input is segmented into arbitrary multi-phone units. The foreign language used in the experiments presented in this paper is the arabic dialect spoken in Iraq.

The paper is organized as follows. In Section 2 we describe our unsupervised ‘word discovery’ (or word segmentation)

¹ See introduction in chapter 5 of [1]

² a measurement we made for French suggests that the phone transcription by a native annotator is approximately 3 to 5 times slower. This measurement is probably very language dependent.

algorithm that is needed to segment the foreign phone sequences into bigger units in order to help the word alignment and also to improve the phone recognition as explained later in the paper. The structure of both *baseline* and *phone-based* speech to speech translation systems and a description of their ASR and MT components are given in Section 3. Section 4 gives the experimental results, where it is shown that using our so-called *phone-based* speech to speech translation approach leads almost to the same performance as the *baseline* approach, while being theoretically applicable to any non written language. Finally, conclusions are drawn in Section 5.

2. UNSUPERVISED WORD DISCOVERY FROM PHONEME SEQUENCES

While phrase-based approaches suggest that it might be possible to directly train translation models between english words and foreign phone sequences, some preliminary experiments have shown that this leads to poor results in practice (between 10% and 15% absolute decrease of BLEU, detailed numbers given in text of *section 4.1*), even when the maximum phrase length is increased. One reason is the low quality of the initial alignments used to build the translation table. To avoid this problem, we propose to preprocess the foreign phone utterances by aggregating phones into longer units before alignment. This is done with the unsupervised word discovery algorithm described in this section.

2.1. Background

Many algorithms for automatic word segmentation use lexicons or pre-segmented text. However, in our hypothesis of non written languages, we are only interested in algorithms that work without supervision and induce, from raw data, a plausible segmentation of a text into words. Such algorithms may be found in NLP literature: [3] [4] and [5] propose approaches for word discovery from raw data; [6] [7] describe unsupervised learning of morphology for highly-inflected languages. Similar algorithms are also found in the computational genomics literature [8].

2.2. Our algorithm

Our algorithm gathers different approaches already applied to the word segmentation problem:

a) use predictability of phonemes: the basic idea here, first suggested by [9], is that the number of distinct phonemes that are possible successors of the preceding string reduces rapidly with the length of that string unless a morph boundary is crossed. A slightly different way to implement this same idea is to compute the mutual information (MI) between all successive phones x and y of an utterance, and to detect a morph boundary when MI reaches a local minimum which is, in the same time, below a certain threshold:

$$MI(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} < \alpha$$

- b) use word boundaries that are already available before (respectively after) phone sequences commonly seen at the beginning (respectively the end) of sentences,
- c) use word frequencies: after a first segmentation, discovered words with high frequency counts are probably real words while words with low counts may result from badly placed word boundaries,
- d) use the strength of Viterbi decoding.

With these ideas, our iterative algorithm is the following:

- 0- *initialization*: perform a first word segmentation of all the foreign corpus using MI criterion only,
- 1- *vocabulary and segment language model training*: build a vocabulary of the 1000 most frequent words found in the last segmented corpus ; put word boundary marks in the unsegmented corpus according to this 1000 word vocabulary and train a n-gram LM from this data,
- 2- *decoding*: for each unsegmented utterance, infer the most likely segmentation (location of segment boundaries) using the language model obtained in step 1,
- 3- *go back to step 1*

It is important to note that in step 2, we make a segment boundary a priori more likely using a bias factor in order to perform a more aggressive segmentation. The reason for this is that a false detection (put an incorrect word segment) may be not too critical for the training of the phrase table, while a false rejection (do not segment multiple words) may freeze some bad sequences before the MT training.

2.3. Performance

For all the experiments reported in this paper, our parallel data is made up of 366k Iraqi utterances and their English translation. To test the potential of our phone-based approach, the words in the Iraqi utterances were replaced by their corresponding phone sequence found in a pronunciation dictionary³. The resulting 366k sequences of phones were sent to our word discovery algorithm, and the segmented output was compared to the reference word segmentation. The performances (word accuracy on all the data, and percentage of real words among the top 1000 most frequent discovered words) obtained for 3 iterations of the above algorithm, are presented in *table 1*.

Iteration	%Word Accuracy	% real words / 1000 most frequent discovered words
1	34.5	74.5
2	51.0	84.1
3	55.2	88.3

Table 1: Performance of our word discovery algorithm

Concerning the performance, it is worth noticing that our reference data is a word segmentation of Arabic, and not a morphological segmentation⁴. A look at some segmentation

³ We are aware that this experimental setup leads to ideal phone sequences which is not the case when working on real 'non written' languages (inconsistency, transcription errors, pronunciation variants). We do not investigate these issues here.

⁴ Arabic words can be decomposed in stem and affixes

outputs shows that our algorithm sometimes perform a morphological segmentation of the data, as a by product, which may partially explain the relatively low word accuracy figures reported. To verify that the algorithm leads to the same performance for other languages, it was also applied to the same amount of English data where the words had been replaced by their phone sequences. The word accuracy obtained after iteration 3 for english is comparable (57.4%).

3. BASELINE VERSUS PHONE-BASED SPEECH TRANSLATION

As already explained in the introduction, both approaches proposed differ in the form of the foreign training data (Iraqi words vs. Iraqi phonemes). Consequently, there is also a difference in the translation process of a foreign sentence.

In the *baseline* approach, the input signal is classically transcribed with an ASR module using a word n-gram language model. The best ASR hypothesis is sent to the decoder that uses phrase-based translation models.

In the *phone-based* approach, the output of the ASR module must be a phone sequence since the phrase-based translation models are learned between foreign phone sequences and their corresponding English text. However, the performance of a phone recognizer than do not use any language model is too weak to provide reliable input to the translation module. One way to improve it (while still respecting the constraint of not using any word information) is to use a vocabulary and n-gram language model trained on the preprocessed foreign data ('words' corresponds then to the phone sequences obtained by the method of *section 2*), as shown in *figure 1*. In that case, the LM used to transcribe foreign speech ($LM(f)$) will not only improve the ASR but will also provide an output consistent with the data used to train the translation model. It is also important to note that we take advantage of the parallel corpus to improve the word discovery itself: the word alignments with high confidence are used to refine the output of the word discovery process and the whole TM is retrained with this new segmented foreign data.

The following subsections shortly describe MT and ASR modules used here which are part of IBM MASTOR.

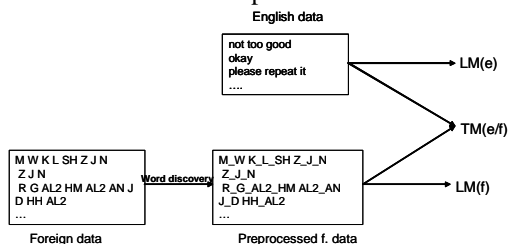


Figure 1: data for LM and TM models in phone-based approach

3.1. Phrase-Based SMT

We employ a memory efficient and fast phrase-based SMT system, named as Folsom [10], that is developed at IBM for speech translation task. This system achieves both memory efficiency and fast speed, which is suitable for real time

speech-to-speech translation on scalable computational platforms. In this new approach, we statically construct a single optimized Finite State Transducer (FST), which is entitled as the Statistical Integrated Phrase Lattice that encodes the entire phrase-level translation model. At runtime, a Viterbi beam decoder is developed to combine the translation model and language model FST's with the input lattice efficiently. The decoding process involves a multiple layer search that can be viewed as an optimized version of on-the-fly or dynamic composition integrated with a Viterbi search procedure. More information about this SMT system is given in [10]. The parallel data is made up of 366k Iraqi utterances and their English translation (2.5M English words and 1.8M colloquial Arabic words).

3.2. ASR

Our ASR for Iraqi Arabic is more precisely described in [11]. While both *baseline* and *phone-based* scenarios described in section 3 may lead to different foreign acoustic models in practice, this point is not discussed here and we used the same acoustic models for all the experiments. The Iraqi arabic acoustic models use 33 phones that essentially correspond to graphemes in the Arabic alphabet. Each phone is modeled with a 3-state left-to-right HMM. The acoustic models are built using 40 dimensional features. The context-dependent model has over 2K leaves and 60K Gaussians. All acoustic models are trained using MPE discriminative training [12]. The training data is about 200 hours of dialectal Iraqi Arabic. The language model uses standard trigrams trained on the foreign part of our 366k sentences parallel data. After removing singleton words, the vocabulary size is 43k for the *baseline* system with "real words"⁵ and 36k for the *phone-based* system with "discovered words".

4. EXPERIMENTS AND RESULTS

4.1. Test set and individual modules evaluation

The test data consists of 3 subsets related to different test scenarios in Iraqi Arabic: *set 1* is made up of 240 speech utterances used for offline S2S evaluation in TRANSTAC in December 2005; *set 2* and *set 3* are made up of respectively 377 and 177 speech utterances which correspond to the live evaluations of TRANSTAC in December 2005 and March 2006. Only *set 1* has multiple (four) english translation references which may partially explain the higher BLEU scores observed on this data set.

The performance of the ASR system on the whole test data (794 utterances) is given in *table 2* for both *baseline* and *phone-based* approaches described in *section 3*. To do a fair comparison between these two, we computed a phone error rate (PER) after replacing each ASR output by its

⁵ While we have shown that a morphological analysis for LM may slightly reduce the word error rate [11], it is not used here and a word is simply defined as a string of characters separated by space.

corresponding phone sequence. We see that the performance of the phone-based system is acceptable compared to the *baseline* system (15.1% against 11.8%) while a simple phone decoder without LM leads to bad performances (42.6% on the same data set). This first result shows that even for non written languages, it is possible to develop an ASR system that gives acceptable PER before translation.

System	%WER	%PER
Baseline	24.8	11.8
Phone-based	X	15.1

Table 2: ASR performance for baseline and phone-based systems

The performance of the translation system from the reference transcripts of the whole test data (794 utterances) is given in table 3 for both *baseline* and *phone-based*⁶ approaches described in section 3. The small differences between both approaches demonstrates that it is possible to train acceptable translation models using a parallel corpus made of foreign phones and English word sequences. Note that the discovery process described in section 2 is necessary since our attempt to bypass it (and directly align foreign phones with words) lead to the following poor BLEU scores: 0.34, 0.26 and 0.24 on test sets 1, 2 and 3 respectively.

System	Test Set 1	Test Set 2	Test Set 3
Baseline	0.49296	0.37316	0.36387
Phone-based	0.47440	0.34559	0.30475

Table 3: Translation-only performance (BLEU scores) for both baseline and phone-based systems

4.2. Complete speech to text evaluation

The performance of the complete speech to text translation of the whole test data (794 utterances) is given in table 4.

System	Test Set 1	Test Set 2	Test Set 3
Baseline	0.43583	0.31007	0.31486
Phone-based	0.45178	0.28640	0.23677
Supervised phone-based	0.46064	0.30414	0.25522

Table 4: Speech-to-text performance (BLEU scores) for baseline, phone-based and supervised phone-based systems

These results confirm that the *phone-based* approach, theoretically applicable to any non-written language, leads to acceptable performance compared to a *baseline* method (the *baseline* is even slightly outperformed on the first test set). To go beyond these BLEU scores, we performed a subjective evaluation (binary decisions) of both *baseline* and *phone-based* output translations: 58% of the *baseline* utterances were judged correctly translated while 54% of the *phone-based* were judged correct (64% of the utterances were judged correctly translated by at least one approach).

4.3. Supervised phone-based approach

The above *phone-based* approach does not use any word information (segmentation of phone sequences is based on

an unsupervised word-discovery process). However, it might be possible in practice to achieve a human supervision of the segmented data. For instance, one may ask a native annotator to indicate, from the unsupervised word discovery process, which words are real and which are non sense. To simulate this, we extracted a list of 2154 real words by discarding the non existing words from the 3000 most frequent discovered words found in our training data. Then, the 366k sequences of phones were re-segmented by aggregating all the phone sequences corresponding to the 2154 real words, and by leaving unchanged the rest of the unknown phone sequences. The preprocessed data obtained was used to retrain the LM and TM as already explained in figure 1. The speech to text translation performance obtained with this *supervised phone-based* approach is given in the last line of table 4 and shows its interest to gradually improve the *phone-based* speech to text performance using native speaker supervision. It is also interesting to note that in this case, the ASR vocabulary is much smaller (since it is made of the 2154 kept words plus the isolated phones) : 2,2k instead of 36k for the *phone-based* approach, while theoretically allowing the same coverage for translation.

5. CONCLUSION

This paper presented a phone-based speech translation approach that bypass the written form of the foreign language and can give almost similar performance to a baseline method. As a by product, this approach may also be interesting to handle foreign proper names (persons, cities, ...) by propagating their recognized phone sequence along the S2S translation process instead of trying to go through their written form which is often unknown or unnormalized.

6. REFERENCES

- [1] T. Schultz, K. Kirchhoff, *Multilingual Speech Processing*, Academic Press, Elsevier, 2006.
- [2] D. Nettle, S. Romaine, *Vanishing Voices, the Extinction of the World's Languages*, Oxford University Press, 2000.
- [3] D. K. Roy, A. Pentland, "Learning words from sights and sounds: a computational model", *Cognitive Science*, Elsevier, 26 (2002).
- [4] C. DeMarcken, *Unsupervised Lang. Acquisition*. PhD, MIT, 1996.
- [5] M. R. Brent, "An efficient, probabilistically sound algorithm for segmentation and word discovery". *Machine Learning*, 34 (1999).
- [6] J. Goldsmith, "Unsupervised learning of the morphology of a natural language". *Computational Linguistics*, 27(01), pp153-198.
- [7] M. Creutz, K. Lagus, "Induction of a Simple Morphology for Highly-Inflecting Languages". In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, Barcelona, 26 July 2005, pages 43-51.
- [8] M. R. Brent, R. Guigo, "Recent advances in gene structure prediction", In *Current Opinion in Structural Biology*, 14 (2004).
- [9] Z. Harris, "From phoneme to morpheme". *Language* 31 (1955).
- [10] B. Zhou, S. Chen and Y. Gao, "Folsom: A Fast and Memory-Efficient Phrase-based Approach to Statistical Machine Translation", accepted to *IEEE/ACL SLT 2006 conference*.
- [11] M. Afify, R. Sarikaya, H-K. J. Kuo, L. Besacier, and Y. Gao "On the use of morphological analysis for dialectal Arabic speech recognition", in *Proc. ICSLP'06*, Pittsburgh, USA, 2006.
- [12] D. Povey, P. Woodland, "Minimum Phone Error and ISmoothing for Improved Discriminative Training," In *Proc. ICASSP 2002*.

⁶ In that case, the word discovery is made on the reference phone transcripts using step 2 (decoding) of our word discovery process (with the segmented LM obtained after iteration 3).