Human Translations Guided Language Discovery for ASR Systems

Sebastian Stüker[1], *Laurent Besacier*[2], *and Alex Waibel*[1]

[1]Institut für Anthropomatik, Universität Karlsruhe (TH), Germany
[2]Laboratory of Informatics of Grenoble (LIG), University J. Fourier, France
stueker@ira.uka.de, laurent.besacier@imag.fr, waibel@ira.uka.de

Abstract

The traditional approach of collecting and annotating the necessary training data is due to economic constraints not feasible for most of the 7,000 languages in the world. At the same time it is of vital interest to have natural language processing systems address practically all of them. Therefore, new, efficient ways of gathering the needed training material have to be found. In this paper we continue our experiments on exploiting the knowledge gained from human simultaneous translations that happen frequently in the real world, in order to discover word units in a new language. We evaluate our approach by measuring the performance of statistical machine translation systems trained on the word units discovered from an oracle phoneme sequence. We improve it then by combining it with a word discovery technique that works without supervision, solely on the unsegmented phoneme sequences.

Index Terms: automatic speech recognition, language discovery, machine translation, under-resourced languages

1. Introduction

Approximately 7,000 living languages exist today. The current edition of Ethnologue lists 7,299, many of which are only spoken by a comparatively small population [1]. In today's globalized world language is seen as one of the last barriers that inhibit cultural exchange and trade [2]. Also, languages are currently dying out at an alarming rate. Linguists estimate that at least half of the living languages will become extinct in this century [3, 4]. While the linguistic diversity might complicate communication among peoples, it is at the same time a keystone to our cultural diversity and therefore worthwhile to maintain [5].

So far, automatic speech recognition (ASR) systems and machine translation (MT) systems have been trained for only a fraction of these languages. Mostly languages with either a large amount of speakers, with a large economic value, or with high political impact have been worked on. However, in order to stop the trend of extinction of languages, it is necessary for technology to address all languages in the world. Also, the political importance of a language might rapidly change. Languages that were uninteresting in the past might suddenly become important, e.g. due to policy changes.

Some linguists estimate, that the vast majority of languages is without a writing system [5]. Omniglot attributes their list of writing systems to only 685 languages [6]. Also, many languages which in theory could be written, are often mostly spoken but very seldomly written. We call these languages *rarely written languages*. For example, Iraqi is a version of Arabic which is mainly spoken, but very seldomly written. So, if one wants to address all languages in the world, one has to prepare for encountering many languages without a writing system or at least no written records.

Also, the training of the models for speech recognition and machine translation systems requires the collection of large amounts of transcribed audio files, sentence parallel corpora in the target languages, and large amounts of monolingual texts in the desired languages and domains.

But, when considering the large amounts of languages in the world and the fact that most of them are only spoken by a relative minority, it becomes clear that the traditional approach of collecting and annotating training data is not feasible for all languages. It is too expensive and time consuming, and requires too much expert knowledge in the languages to be addressed.

Therefore, we have started to address the problem of automatically exploring a new language for which no ASR or MT systems have been created so far. Our focus has been on a particular scenario where a human translator is available. When communication with speakers of a less resourced language, maybe even one without a written representation, becomes necessary, it is often achieved with the help of bilingual human translators, a very costly resource. Our goal is now to exploit the translations of the human interpreter, in order to gather the material needed for training ASR and translation systems.

Since many of the less resourced languages might not have a writing system or no expert competent in the existing writing system is available, we assume that we can only work with a phonetic transcription of that language. Such a transcription can be obtained manually by skilled phoneticians. Alternatively work is going on to automatically obtain such a description in any kind of language [7, 8].

In [9] we proposed a method to automatically find word like units in phonetic transcriptions of speech that are only segmented at a sentence level. As quality measure of the found word units we used the translation performance of a statistical Iraqi to English translation system that was trained on the discovered, Iraqi word units and regular English words. The method for finding the Iraqi words works on the phoneme sequence alone without considering the English translations used for training the statistical translation system

In [10] we examined the feasibility of automatically learning word units in the unknown language and their pronunciation by aligning the English word sequences in the training corpus with the phonetic transcription of the minority language. This time we worked on an English-Spanish corpus, pretending that Spanish was the new language in which we wanted to discover word units. In these first experiments we did not work with a full translation system. Instead, we only measured the number of discovered words that actually correspond to real Spanish words by giving precision and recall.

In this paper we have extended our experiment from [10] by introducing a method for applying the words, found through the alignment, to sets of data for which no parallel data is available. We measure the quality of the found words by the performance of a statistical translation system trained on these words units and compare it against the performance reached by the word discovery procedure from [9]. We then combine the two methods in order to further improve the translation quality obtainable with the found words.

2. Set-up and Database

Our experiments were conducted with the help of the English portion of the Basic Travel Expression Corpus (BTEC) [11] and a Spanish translation of it. BTEC consists of travel expressions taken from phrase books in order to cover every potential subject in travel conversations. Our version of BTEC with the corresponding Spanish translation of it consists of 146K parallel sentences. The size of the English vocabulary is 12K while that of the Spanish one is 17K.

In our experiments English plays the role of the well-studied language while Spanish takes the role of the under-resourced language about which little to nothing is known and we assume that we can only obtain a phonetic transcription of the Spanish sentences instead of a word transcription. In reality Spanish is of course a well known language with existing resources and systems. However, by pretending that it is unknown to us, we can simulate our approaches on existing data and can easily evaluate them.

For our exploratory experiments we use a perfect phoneme transcription of the Spanish sentences which we obtained by transforming the words in the Spanish corpus with the help of a dictionary that was generated by a rule based system. As mentioned above, on real-data, this phonetic transcript could either be obtained manually or with automatic methods. Especially the latter ones will introduce errors in the transcripts which we want to exclude for now from our experiments; especially since the development of multilingual, automatic phoneme transcriptions is still a progressing area.

We divided the corpus into three sets: a training set containing 142,810 sentence pairs, a development set with 2,000 sentence pairs, and a test set with 2,000 sentences. The development set is used for parameter tuning for the translation system, while the performance of the resulting translation systems is measured on the test set. All three sets are completely disjoint and do not contain any doublets.

3. Monolingual Word Discovery

Our first algorithm works without supervision and finds a plausible segmentation of a text into words by solely working on the Spanish phonemes. Such algorithms may be found in *natural language processing* (NLP) literature: [12], [13], and [14] propose approaches for word discovery from raw data; [15] and [16] describe unsupervised learning of morphology for highly-inflected languages. Similar algorithms are also found in the computational genomics literature [17].

Our algorithm gathers different approaches already applied to the word segmentation problem: a) use predictability of phonemes: the basic idea here, first suggested by [18], 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 idea is to compute the *mutual information* (MI) between all successive phonemes of an utterance, and to detect a morph boundary when MI reaches a local minimum which is, at the same time, below a certain threshold: 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 consists of three steps: 1) initialization: perform a first word segmentation of the foreign training corpus using the MI criterion only, 2) 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, 3) decoding: for each unsegmented utterance, infer the most likely segmentation (location of segment boundaries) using the language model obtained in step 2, 4) go back to step 1.

More details on this algorithm can be found in [9].

4. Word Discovery from Parallel Data

In our second approach we do not only work on the Spanish phonemic transcripts but include the parallel, English sentences as well into the word discovery procedure. We segment the Spanish phoneme string into appropriate word units by establishing word-to-phoneme alignments between the individual English words and chunks from the phoneme sequence. The alignments are then transformed into a dictionary by associating the aligned, consecutive phoneme blocks with words, depending to which English words they have been aligned to. The resulting dictionary of words and their phoneme sequence, is then filtered in order to eliminate artifacts and wrongly aligned words. The phonemes of the Spanish training, development, and test set are then replaced by the words in the extracted dictionary.

4.1. Word Alignment

The science of establishing word-to-word alignments for bilingual sentences has been well studied in the field of Machine Translation. The alignment between a given source string with J words $s_1^J = s_1, s_2, ..., s_J$ and a target string with I words $t_1^I = t_1, t_2, ..., t_I$ is defined as a subset of the Cartesian product between the word positions of the two strings [19, 20]:

$$A \subseteq \{(i, j) : j = 1, ...J; i = 1, ..., I\} \quad (1)$$

Usually the alignments are constrained in such a way that each source word is assigned exactly one target word; so for every word position j in the source sentence a word position $i = a_j$ in the target sentence is assigned and we can write the alignments as $a_1^J = a_1, ..., a_J$.

One solution to automatically finding such alignments between two sentences now is the use of statistical alignment models and statistical translation models from statistical machine translation (SMT)[19]. One part of SMT tries to model the translation probability $P(s_1^J|t_1^I)$ which describes the relationship between a source language string s_1^J and a target language string t_1^I. Now, given the alignment a_1^J between s_1^J and t_1^I a statistical alignment model is defined as $P(s_1^J, a_1^J|t_1^I)$, and $P(s_1^J|t_1^I)$ can be expressed as

$$P(s_1^J|t_1^I) = \sum_{a_1^J} P(s_1^J, a_1^J|t_1^I) \quad (2)$$

The statistical models in general depend on a set of parameters Θ: $P(s_1^J, a_1^J | t_1^I) = P_\Theta(s_1^J, a_1^J | t_1^I)$. The best parameters $\bar{\Theta}$ are found on a set S of parallel training sentences, in such a way that they maximize the probability of the training set. One way to do this is to use *expectation maximization* (EM) training which in general will only find a local maximum for $\bar{\Theta}$. Given a sentence pair (s_1^J, t_1^I) the best alignment, that is the most probable alignment, between the two sentences can be found with the help of the trained parameters:

$$\bar{a}_1^J = \operatorname*{argmax}_{a_1^j} P_{\bar{\Theta}}(s_1^J, a_1^J | t_1^I) \tag{3}$$

For our experiments we use the IBM-4 model to generate the sentence alignments [20].

5. Experiments

5.1. Monolingual Word Discovery

We applied our algorithm from Section 3 to the Spanish training phoneme sequence on a per sentence level. The initial MI threshold was 1.0. 3 iterations of language model training and segmentation were performed. From the word units found on the training data we then extracted a dictionary with occurrence counts for the words. Using this dictionary we substituted the phonemes in the training, development, and test set by recursively substituting the longest matching phoneme sequence. It is important to note that in step 2 of the algorithm, 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.

5.2. Word Discovery from Parallel Data

In the approach described in Section 4 we want to assign every English word a sequence of Spanish phonemes. For finding the word alignments we used the GIZA++ [21] toolkit and the Pharaoh training script [22]. One result of the GIZA++ training besides the learned translation models is a word alignment for the sentences in the training set. Since the alignments have the restriction that each source word is assigned exactly one target word, English is the target language and Spanish the source language. In order to get an impression of the alignments found by the training, Figure 1 shows three sample alignments between the English words (top) and the Spanish phonemes (middle). Below the Spanish phonemes the figure shows the Spanish word transcription together with the word to phoneme mapping as given by our dictionary. The alignment a) in this figure is an example for a perfect alignment in a rather simple case, where the number of English words matches the number of Spanish words. b) is an example of a more complex alignment where the English word ‘please’ needs to be aligned to two Spanish words. Again the alignment found is correct. c) shows an example of an even more complicated alignment. Here the alignment also needs to do a word reordering, the words ‘hot’ and ‘milk’ need to be swapped. And the English words ‘I’d’ and ‘like’ both need to be mapped to the Spanish word ‘querria’. While the swap of ‘hot’ and ‘milk’ is done correctly, the alignment found for ‘I’d’ and ‘like’ is clearly wrong. Due to its constraints the IBM-4 model cannot find the correct alignment.

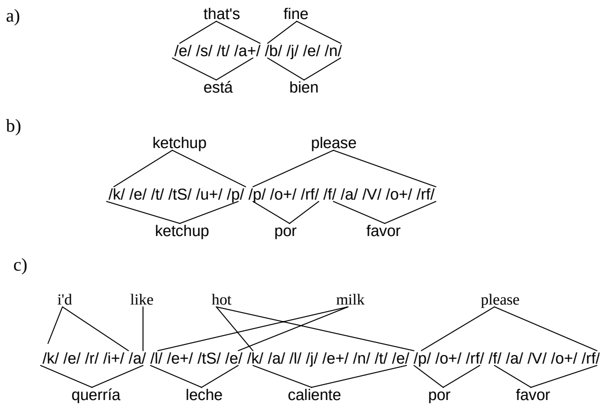
a)
that's
fine
/e/ /s/ /t/ /a+/ /b/ /j/ /e/ /n/
está
bien
b)
ketchup
please
/k/ /e/ /t/ /tS/ /u+/ /p/ /p/ /o+/ /rf/ /f/ /a/ /V/ /o+/ /rf/
ketchup
por
favor
c)
i'd
like
hot
milk
please
/k/ /e/ /r/ /i+/ /a/ /l/ /e+/ /tS/ /e/ /k/ /a/ /l/ /j/ /e+/ /n/ /t/ /e/ /p/ /o+/ /rf/ /f/ /a/ /V/ /o+/ /rf/
querría
leche
caliente
por
favor

Figure 1: Samples of alignments found by GIZA++

5.2.1. Dictionary Extraction and Corpus Preparation

From the found alignments it is now easily possible to extract dictionary entries. Every English word that is aligned to Spanish phonemes is a potential entry in the Spanish dictionary, with the English word serving as a generic word id in the Spanish dictionary. Different English words that were mapped to the same phoneme sequence were not combined into one word, so that homophones were generated. One special case, when extracting the words, needs to be considered. It can happen that an English word is aligned to a phoneme sequence that is not continuous in its phonemes’ positions, but that has got holes or reorderings in its sequence. These sequences have to be split into its continuous subsequences, each subsequence corresponding to one Spanish word. Each subsequence then receives its own word identifier based on the English word to which it was aligned.

In a second step the resulting dictionary is filtered. Pronunciation variants to a word that occur less than 100 times in the training text are removed from the dictionary. This step is taken in order to eliminate pronunciation variants that were created due to erroneous word-to-phoneme alignments. The dictionary constructed in this way contains 15K words. 3,172 words in the original Spanish dictionary have an exact phonetic match in the dictionary constructed this way.

Using the extracted dictionary the phoneme sequences in the training, development, and test set were replaced by the words in the dictionary. Replacement was done by recursively replacing the longest matching phoneme sequence in a sentence with the corresponding word from the dictionary. In case of multiple, matching words the most frequent one was chosen.

5.3. Combination of the Approaches

We now combined the two approaches in order to level on the complementary information that they use for the word segmentation. We did this by taking the most frequent words from the initial mutual information segmentation from the monolingual approach and replacing these found words in the Spanish corpus. After that the regular segmentation procedure from Section 4 was performed.

5.4. Evaluation

Using the corpora from the resulting word discovery procedures Spanish-to-English translation systems were trained with them using the Moses toolkit [23]. We performed the standard training and decoding procedure as described on the Moses homepage. We also trained a translation system with the original Spanish, word based corpus which serves as our gold standard, and whose performance provides us with an upper bound for the performance of the other systems.

The results of the evaluation on the English test set, when translating from Spanish, are summarized in Table 1. All results are in BLEU. Using the original words, we achieve a BLEU score of 0.56. When we use the monolingual word discovery procedure described in Section 3 instead, we loose significantly in translation performance and only reach a score of 0.39. By exploiting the available parallel English data as described in Section 4 we achieve a significantly better BLEU score of 0.50, losing only six BLEU points to the gold standard.

When we combine the two approaches as described in Section 5.3, we can improve on that performance by one BLEU point reaching a score of 0.51. The number of words to take from the initial MI segmentation was empirically determined on the development set and set to 40.

Table 1: Results in BLEU Score of the Spanish-to-English translation with the different word discovery approaches

Word Segmentation Approach	BLEU
Gold Standard	0.56
Monolingual	0.39
Parallel Data	0.50
Combination	0.51

6. Conclusion

In this paper we evaluated a technique for discovering words in a new, unknown language which uses parallel, bilingual data of phoneme sequences and words—as it can, for example be obtained by observing human interpreters. As evaluation metric we used the translation performances of a statistical machine translation system that was trained on the automatically discovered word units. The word units were discovered from an unsegmented phoneme sequence of the training material in the training language. For our exploratory experiments we worked on a perfect oracle phoneme sequence which we created by applying a pronunciation dictionary to the original word sequence. We compared this approach with another technique which works without any supervision and only operates on the phoneme string in the new language without considering the parallel data in the other training language. By combining the two approaches we could show a gain of one BLEU point. The approaches presented in this paper are especially useful for under-resourced languages for which it is not possible to invest large amounts of money and man power for system development.

7. Acknowledgments

This work was in part realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

[1] R. G. Gordon, Jr. and B. F. Grimes, Eds., *Ethnologue: Languages of the World*. Dallas, Texas, USA: SIL International, 2005.

[2] V. Steinbiss, "Human language technologies for europe," Work comissioned by ITC-irst, Trento, Italy to Accipio Consulting, Aachen, Germany, April 2006.

[3] T. Janson, *Speak — A Short History of Languages*. Oxford, UK: Oxford University Press, 2002.

[4] D. Crystal, *Language Death*. Cambridge, UK: Cambridge University Press, 2000.

[5] D. Nettle and S. Romaine, *Vanishing Voices*. New York, NY, USA: Oxford University Press Inc., 2000.

[6] "Omniglot writing systems and languages of the world," http:\\www.omniglot.com.

[7] J. Köhler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proceedings of the Fourth International Conference on Spoken Language Processing*. Philadelphia, PA, USA: ISCA, October 1996, pp. 2195–2198.

[8] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, August 2001.

[9] L. Besacier, B. Zhou, and Y. Gao, "Towards speech translation of non written languages," in *Spoken Language Technology Workshop, 2006*. Aruba: IEEE, December 2006, pp. 222–225.

[10] S. Stüker, "Towards human translations guided language discovery for asr systems," in *Proceedings of the First International Workshop on Spoken Languages Technologies for Under-resourced languages (SLTU)*, Hanoi, Vietnam, May 2008.

[11] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH'03*. Geneve, Switzerland: ISCA, September 2003, pp. 381–384.

[12] D. K. Roy and A. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, January-February 2002.

[13] C. DeMarcken, "Unsupervised language acquisition," Ph.D. dissertation, Unsupervised Language Acquisition, September 1996.

[14] M. R. Brent, "An efficient, probabilistically sound algorithm for segmentation and word discovery," *Machine Learning*, vol. 34, no. 1-3, pp. 71–105, February 1996.

[15] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Computational Linguistics*, vol. 27, no. 2, pp. 153–198, June 2001.

[16] M. Creutz and 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, July 2005, pp. 43–51.

[17] M. R. Brent, "Recent advances in gene structure prediction," *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 264–272, May 2004.

[18] Z. S. Harris, "From phoneme to morpheme," *Language*, vol. 31, no. 2, pp. 190–222, April-June 1955.

[19] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003.

[20] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[21] F. J. Och and H. Ney, "Improved statistical alignment models," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. Hongkong, China: Association for Computational Linguistics Morristown, NJ, USA, October 2000, pp. 440–447.

[22] P. Koehn, "Pharaoh: a beam search decoder for phrase-based statistical machine translation models," 2004, http:\\www.isi.edu\licensed-sw\pharaoh\.

[23] "Moses, a factored phrase-based beam-search decoder for machine translation," http:\\www.statmt.org\moses\.