TOWARDS HUMAN TRANSLATIONS GUIDED LANGUAGE DISCOVERY FOR ASR SYSTEMS

Sebastian Stüker and Alex Waibel

Institut für Theoretische Informatik
Universität Karlsruhe (TH)
Karlsruhe, Germany
stueker@ira.uka.de, waibel@ira.uka.de

ABSTRACT

Natural language processing systems, e.g for Automatic Speech Recognition (ASR) or Machine Translation (MT), have been studied only for a fraction of the approx. 7000 languages that exist in today's world, the majority of which have only comparatively few speakers and few resources. The traditional approach of collecting and annotating the necessary training data is due to economic constraints not feasible for most of them. At the same time it is of vital interest to have NLP systems address practically all languages in the world. New, efficient ways of gathering the needed training material have to be found. In this paper we propose a new technique of collecting such data by exploiting the knowledge gained from Human simultaneous translations that happen frequently in the real world. To show the feasibility of our approach we present first experiments towards constructing a pronunciation dictionary from the data gained.

Index Terms— Automatic Speech Recognition, Language Discovery, Machine Translation, Under-Resourced Languages

1. INTRODUCTION

1.1. The Traditional Way to Acquire Training Data

Training large vocabulary continuous speech recognition (LVCSR) systems requires a number resources in the targeted language. For training the acoustic model of a recognition system large amounts of transcribed audio recordings of speech are needed. The training of the language model requires large amounts of written text in the targeted language. When using phoneme based acoustic models, a pronunciation dictionary is needed that maps the written representation of a word to the sequence of its phonemes when being spoken.

Approximately 7,000 languages exist today, the current edition of Ethnologue [1] lists 7,299. So far, automatic speech recognition (ASR) systems and machine translation (MT) systems have been trained for only a fraction of these languages. Languages addressed so far are either languages with a large amount of speakers, with a large economic value, or with high political impact. Thus, for these languages the resources needed for training speech recognition and translation systems were either available, or it was feasible to invest the time, money, and man power needed to create the required resources. Usually the resources are generated by collecting existing texts in the target language and by manually annotating speech recordings in the corresponding language. These speech recordings can be either already existing ones, e.g. Broadcast News, or can be generated by performing dedicated data collection efforts for the development of the ASR systems. In general, the acquisition of the necessary resources in this way is quite expensive.

However, for the majority of the 7,000 or so languages in the world this traditional approach for gathering or generating the required resources is not feasible. Due to their often less prevalent position — when using the indicators "number of speakers, "economic value, and "political relevance — it is not economical to invest the same amount of money into generating the training resources, as for the languages studied so far. Further, due to their less prevalent position, the underlying resources from which the training material is often generated, e.g. TV and radio broadcasts or press agency releases, are not available; thus making it even more expensive and time consuming to generate the training material the old-fashioned way. Also, many of these under-resourced languages do not meet certain conditions that were given in the well resourced languages studied so far. For example, many of the under-resourced languages do not have a writing system. Or, if a writing system exists, written resources do not exist, because a significantly different language that is only related to the language in question is used for written communications. For example, the language Iraqi is frequently used for oral communication, but seldomly written. Modern standard Arabic is used instead for written records.

1.2. The Need to Address Less Prevalent Languages

Despite these factors, that at a first glance might make a language appear less important, and thus unnecessary as tar-

The authors would like to thank Muntsin Kolss and Matthias Paulik for their help with BTEC, the GIZA++ toolkit, and questions regarding the field of SMT in general. The authors would also like to thank the reviewers for their helpful remarks.

get for natural language processing (NLP) technologies, good reasons exist for developing ASR or MT systems for literally all languages in the world. First, the diversity of languages in the world is the basis of the rich cultural diversity in the world. However in today's globalized world, languages are frequently disappearing. In [2] Janson estimates that in a few generations at least 1,000 of todays languages will have disappeared and that, if the trend holds, in as little as one hundred years half of todays languages will be extinct. With this loss of languages comes a loss in cultural diversity which needs to be prevented. The ongoing extinction of many languages is in part caused by a switch to more prevalent languages that might give their speakers an economic advantage. The lack of NLP systems for this languages only accelerates their extinction while on the other side NLP could help to stop this trend by making the less prevalent languages more attractive to their original speakers.

A second reason why NLP techniques should be available for all languages is that the political impact of a language can be very volatile. In today's globalized world, language is one of the few remaining barriers that hinder human-to-human interaction. Events such as armed conflicts or natural disasters might make it important to be able to communicate with speakers of a less-prevalent language, e.g. for humanitarian workers in a disaster area. Often the people that one need to communicate with in such a scenario only speak their own language that is unknown to the outsider, e.g. a foreign doctor trying to help. For these cases Human translators are often not available in necessary numbers and in a timely manner. Here, readily available NLP technology such as speech translation systems can be highly beneficial. NLP technology might be far from being perfect, but when being faced with the alternative of having no translation system at all for an unknown language in an emergency situation, the imperfect system will be of great use. Therefore, NLP needs to be developed especially for under-resourced languages.

1.3. Goals of this Paper

In this paper we will present our first experiments in developing techniques for exploring a new language that has not been addressed by NLP so far, and for acquiring the necessary training resources for building ASR systems in that new language. We focus 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. For example, English speaking doctors in a remote disaster area might communicate with their patients with the help of a Human translator. 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. In our experiments we examine the feasibility of automatically learning word units in the unknown language and their pronunciation by aligning the English word sequences, that are being translated by the Human interpreter, with the phonetic output from the translator's speech. We assume that we have no knowledge about a potential writing system in the target language nor about possible word units. Thus we are only able to work with the phonetic representation of the interpreter's speech. Besacier et.al. proposed in [3] to train speech translation systems on data that contains English words on the one side and phonemes on the other side, and conducted experiments on English words and Iraqi phonemes. In order to achieve good translation performance [3] first ran a word discovery algorithm on the Iraqi phonemes without considering the corresponding English word sequence and then trained the translation system on the discovered word like units. In our experiments we perform the word discovery by utilizing the knowledge that can be gained from automatically aligning the English word sequences with the Iraqi phoneme sequences. We feel that the English word sequence which is known to correspond to the Iraqi phonemes should give additional information that can be used for the word discovery.

2. EXPERIMENTAL SETUP

The goal of our experiments is to automatically exploit the data that is generated in the Human interpreter scenario described above. We assume that one of the languages involved is a well known language that has been examined already for NLP, meaning that for example ASR systems for this language exist. English is such a language that is often used in scenarios as described here. For the other language it is only assumed that a phonetic transcript of the words articulated by the translator is available. In a real-world application scenario this transcript has to be obtained in an automatic way by a language independent phoneme recognition system. The construction of such systems is an area of research by itself (e.g. [4], [5]). For the experiments in this paper we chose to work with a reference phoneme transcription of the target speech, instead of automatic ones. In this way we want to exclude effects introduced by errors in the phoneme recognition of the target language and concentrate on the techniques for exploiting the parallel data.

2.1. Word Alignment

In order to segment the phoneme string of the target language into appropriate word units we propose to exploit the original English speech by establishing word-to-phoneme alignments between the individual English words and chunks from the phoneme sequence. 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 [6], [7]:

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

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)[6]. 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:

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

Different models, with different sets of parameters exist in literature, such as HMM models [8] and the IBM 1-5 models [7]. For our experiments we use the IBM-4 model to generate the sentence alignments.

2.2. Alignment Error Rate

For assessing the quality of the found alignments between two sentences, [6] defines the alignment error rate (AER). For calculating the AER a set of manually annotated reference alignments is created. Due to the complexity and ambiguity of creating a reference alignment, the alignments a_j are labeled as either belonging to sure (S) alignments or possible (P) alignments, which are used for ambiguous alignments. Every sure alignment is also considered to be a possible alignment ($S \subseteq P$). The quality of the alignment found is then measured by appropriately defined precision and recall measures:

$$recall = \frac{|A \cap S|}{|S|}, precision = \frac{|A \cap P|}{|A|} \quad (4)$$

Thus a recall error only occurs if a sure alignment has not been found, while a precision error occurs if a found alignment is not even possible. The alignment error rate (AER) is derived from the well known F-measure:

$$AER(S, P; A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \quad (5)$$

	Spanish phonemes	Spanish words
precision	83.5%	88.8%
recall	66.9%	75.3%
AER	25.4%	18.1%

Table 1. Precision, recall, and AER for the alignments between English words and Spanish phonemes\words

3. EXPERIMENTS

3.1. Data

Our experiments were conducted with the help of the English portion of the Basic Travel Expression Corpus (BTEC) [9] 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 155K parallel sentences. The size of the English vocabulary is 12K while that of the Spanish one is 20K.

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 As mentioned above, 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.

3.2. Word-to-Phoneme Alignment

In the word to phoneme alignment we want to assign every English word a sequence of Spanish phonemes. For finding the word alignments we used the GIZA++ [10] toolkit and the Pharaoh training script [11]. Before applying the GIZA++ training, sentence pairs from the training corpus were removed that were longer than 50 words or phonemes respectively and that exceeded a sentence length ration of 9-1. 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 have a baseline number for the error rate of the alignments from the training, we also performed the IBM-4 model training for the word based Spanish corpus, instead of the phoneme based one. Table 1 shows the precision, recall, and alignment error rate for the training on the bilingual corpus using Spanish phonemes and the bilingual corpus using Spanish words as a comparison. The alignment error rate for the alignment between the English words and the Spanish phonemes is, as would be expected, higher than for the alignment with the Spanish words. This is due to the more complex task of aligning words with phonemes, instead of words. However, the numbers also show that the task is feasible and can be done with the existing alignment techniques. 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.

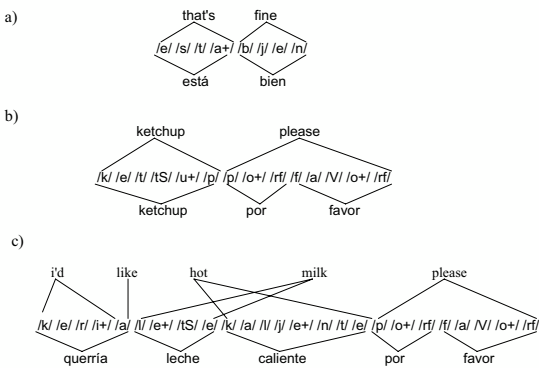

Fig. 1. Samples of alignments found by GIZA++

3.3. Dictionary Extraction

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 can be combined into one word if desired. Otherwise homophones will be potentially 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 can have 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. The, in this way constructed dictionary, contains 16K words. 5,400 words in the constructed dictionary have an exact, phonetic match in the Spanish dictionary from which the phoneme transcription for the model training was constructed.

4. SUMMARY AND FUTURE WORK

In this paper we proposed a new technique for efficiently acquiring the language resources necessary for training ASR systems for new languages about which little or nothing is known. The technique presented in this paper exploits the data generated by human translators, as it is frequently generated in real-life, in an efficient way, making use of the parallel, bilingual nature of the data. The approach is especially useful for under-resourced languages for which it is not possible to invest large amounts of money and dedicated man power for system development. We have demonstrated the feasibility of our approach by automatically constructing a pronunciation dictionary from the acquired parallel data. The construction process is even suitable for languages without a writing system. Future work will focus on improving the algorithms for extracting the dictionary from the allocated data and will extend the data exploitation to building all components of a full-fledged speech-to-speech translation system.

5. REFERENCES

[1] R. G. Gordon Jr., Ed., *Ethnologue, Languages of the World*, SIL International, fifteenth edition, 2005.

[2] T. Janson, *Speak – A Short History of Languages*, Oxford University Press, 2002.

[3] Laurent Besacier, Bowen Zhou, and Yuqing Gao, "Towards Speech Translation of Non Written Languages," in *IEEE/ACL SLT Workshop*, Aruba, December 2006.

[4] Joachim Köhler, "Multi-Lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities Of Sounds," in *ICSLP*, Philadelphia, PA, USA, October 1996.

[5] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, August 2001.

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

[7] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.

[8] Stephan Vogel, Hermann Ney, and Christopher Tillmann, "HMM-based word alignment in statistical translation," in *COLING*, Copenhagen, Denmark, August 1996.

[9] Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto, "Creating corpora for speech-to-speech translation," in *Interspeech*, Geneve, Switzerland, September 2003.

[10] F. J. Och and H. Ney, "Improved statistical alignment models," Hongkong, China, October 2000, pp. 440–447.

[11] Philipp Koehn, "PHARAOH: a beam search decoder for phrase-based statistical machine translation models," 2004, http://www.isi.edu/licensed-sw/pharaoh/.