WORD SEGMENTATION THROUGH CROSS-LINGUAL WORD-TO-PHONEME ALIGNMENT

Felix Stahlberg[*], *Tim Schlippe*[*], *Stephan Vogel*[†], *Tanja Schultz*[*]

[*] Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany
[†] Qatar Computing Research Institute, Qatar Foundation, Qatar

ABSTRACT

We present our new alignment model *Model 3P* for cross-lingual word-to-phoneme alignment, and show that unsupervised learning of word segmentation is more accurate when information of another language is used. Word segmentation with cross-lingual information is highly relevant to bootstrap pronunciation dictionaries from audio data for Automatic Speech Recognition, bypass the written form in Speech-to-Speech Translation or build the vocabulary of an unseen language, particularly in the context of under-resourced languages. Using *Model 3P* for the alignment between English words and Spanish phonemes outperforms a state-of-the-art monolingual word segmentation approach [1] on the BTEC corpus [2] by up to 42% absolute in F-Score on the phoneme level and a GIZA++ alignment based on IBM Model 3 by up to 17%.

Index Terms— alignment model, word segmentation, under-resourced language, speech-to-speech translation

1. INTRODUCTION

There are over 6,900 living languages and dialects in the world [3]. Automatic Speech Recognition (ASR) and Machine Translation (MT) systems exist only for a few of them due to the large amount of training data needed to train them. Transcribed speech resources, large amounts of text for language modeling, pronunciation dictionaries, and parallel text corpora are of great importance to create speech processing systems. However, languages with few linguistic resources may suddenly appear in the focus of interest. For instance, in the scope of international relief operations, porting language technology rapidly and economically to new unseen and under-resourced languages is in particular suitable. Another challenge is that a lot of the world's languages and dialects do not have a written script despite their widespread use for oral communication [4, 5], whereas language technology generally requires a written script nowadays.

In this paper, we take first steps towards gathering training data for ASR and MT systems for an unseen and under-resourced target language rapidly and at low cost: We segment phoneme sequences into word units using information from another language. These word units can be used to bootstrap and enrich pronunciation dictionaries from audio data for ASR or build the vocabulary of an unseen language. They can also be used to bypass the written form in those systems in order to save costs for manual transcriptions and tackle non-written languages.

Our method benefits from the fact that written sentences are available in several economically viable languages such as English. We assume that a speaker is available who understands English and who speaks translations of the English sentences in his or her mother tongue. This is not a strong assumption, since human simultaneous translations happen frequently in the real world. Our goal is to exploit the phonetic output from the audio recordings of such human translators, so that the following scenario comes within reach (Figure 1):

1) We recognize the spoken translations with a language independent phoneme recognizer. 2) We build an *alignment* between words in the written English sentence and phonemes in the corresponding recognized phoneme sequence in the target language. 3) Using this cross-lingual alignment, we segment the phoneme sequence into word units. 4a) The word segmentation induces phonetic transcriptions of words in the target language, which are used in a pronunciation dictionary for ASR systems. 4b) The segmented phoneme sequence is replaced by a sequence of target word tokens. This results in a parallel corpus on the word level, which serves as training data for a Statistical MT (SMT) system as described in [6].

Since we operate only at the phoneme level on the target side, we implicitly introduce an artificial writing system, where the words are represented by their pronunciations.

This paper presents the first steps towards this scenario by mainly focusing on *steps 2* and *3*. We present a new alignment model *Model 3P*, which is suitable for cross-lingually aligning words to phonemes. Furthermore, we investigate word segmentation and alignment quality with different phoneme error rates. In the next section, we present monolingual and cross-lingual methods of other researchers for word segmentation. Section 3 gives an overview of the functionality of the alignment model IBM Model 3. In Section 4, we introduce *Model 3P*. Our experimental setup is described in Section 5.

We thank the Qatar Foundation for funding a research visit of the first author at the Qatar Computing Research Institute (QCRI).

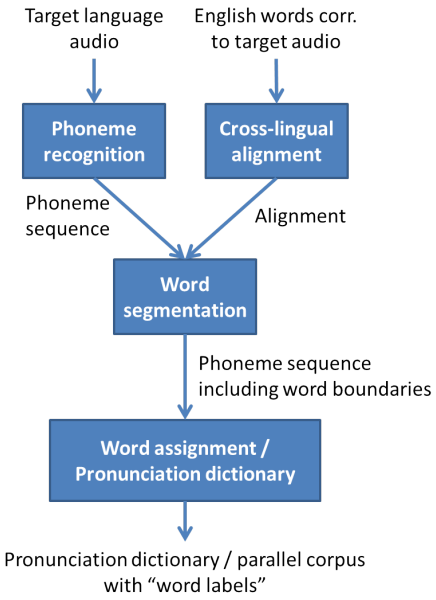


Fig. 1. Scenario.

We present our results in Section 6. In Section 7, we conclude our work and describe our further steps.

2. RELATED WORK

The following monolingual methods for word segmentation have been used in the past: First, Minimal Description Length analysis [7, 8] approximates the optimal compression of a (phoneme-)string (corresponding to its Kolmogorov complexity). Assuming that a word sequence of a language is the optimal compression of the corresponding phoneme sequence, the data segmentation induced by such compression methods is taken as the word segmentation. The second approach uses adaptor grammars, which are context-free grammars that learn new rules from the training data [9]. Since recent studies underline the feasibility of applying adaptor grammars to the word segmentation problem [1], we representatively use them in Section 6 for a monolingual word segmentation method.

In our approach, we use the information of a parallel corpus between word sequences in the source language and phoneme sequences in the target language similar to [6]. In an oracle experiment, they replace the words in the target language with their pronunciations and remove word boundary markers. For segmenting these phoneme sequences into words, however, they run a monolingual unsupervised algorithm in contrast to using cross-lingual word-to-phoneme alignment. They use the resulting word sequences in the training process of an MT system and show that even with low word accuracy of their word segmentation algorithm (55.2%), vocabulary extraction efforts are applicable to MT.

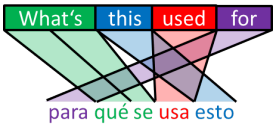


Fig. 2. Word alignment.

The authors in [10] use the word-to-word aligner GIZA++ [11] to align English word sequences to Spanish phoneme sequences to extract training data for language technology from human simultaneously spoken translations. In this paper, we show that even if the found word-to-phoneme alignments in [10] have acceptable quality on correct phonetic transcriptions, the word segmentation precision is not significantly higher than in a monolingual approach when phoneme recognition errors are more common. Therefore, we propose a new alignment model for word-to-phoneme alignment and achieve significantly higher word segmentation and alignment quality on correct phoneme sequences as well as with simulated phoneme recognition errors.

3. IBM MODEL 3

As our approach is highly inspired by ideas originating from SMT, we briefly discuss those in this section. The central data structure in SMT is the word alignment, which identifies word pairs in parallel corpora that are translations of one another. For instance, the alignment in Fig. 2 indicates that the Spanish word `esto` is a possible translation of the English word `this`. Various statistical models for estimating the probability of such alignments exist in the literature, such as the HMM model [12], the IBM Model hierarchy 1-5 [13], and their variations [14, 11]. They differ in the set of parameters, forms of restrictions or deficiency. GIZA++ [11] is an implementation of the IBM Models and the HMM model widely used in SMT for automatically finding word-to-word alignments.

Our proposed alignment model (*Model 3P*) is an extension of the IBM Model 3 [13]. The parameters of the latter model are composed of a set of *fertility probabilities* $n(\cdot|\cdot)$, p_0, p_1, a set of *translation probabilities* $t(\cdot|\cdot)$, and a set of *distortion probabilities* $d(\cdot|\cdot)$. According to IBM Model 3, the following generative process produces the target language sentence f from a source language sentence e with length l [15].

1. For each source word e_i indexed by $i = 1, 2, ..., l$, choose the fertility ϕ_i with probability $n(\phi_i|e_i)$.
2. Choose the number ϕ_0 of "spurious" target words to be generated from $e_0 =$ `NULL`, using probability p_1 and the sum of fertilities from *step 1*.
3. Let $m = \sum_{i=0}^{l} \phi_i$.
4. For each $i = 0, 1, 2, ..., l$, and each $k = 1, 2, ..., \phi_i$, choose a target word τ_{ik} with probability $t(\tau_{ik}|e_i)$.

5. For each $i = 1, 2, ..., l$, and each $k = 1, 2, ..., \phi_i$, choose target position π_{ik} with probability $d(\pi_{ik}|i, l, m)$.
6. For each $k = 1, 2, ..., \phi_0$, choose a position π_{0k} from the $\phi_0 - k + 1$ remaining vacant positions in $1, 2, ..., m$, for a total probability of $1/\phi_0!$.
7. Output the target sentence with words τ_{ik} in positions π_{ik} $(0 \leq i \leq l, 1 \leq k \leq \phi_i)$.

Figure 3 illustrates the generation of the Spanish sentence *Para qué se usa esto* from the English sentence *What's this used for*. Equation 1 states the process as a general formula:

$$P(a, f|e) = \binom{m - \phi_0}{\phi_0} \cdot {p_0}^{m-2\phi_0} \cdot {p_1}^{\phi_0} \cdot \prod_{i=1}^{l} n(\phi_i|e_i) \cdot \prod_{j=1}^{m} t(f_j|e_{a_j}) \cdot \prod_{j:a_j \neq 0}^{m} d(j|a_j, l, m) \prod_{i=1}^{l} \phi_i! \quad (1)$$

The alignment a is represented as a vector of integers, in which a_i stores the position of the source word connected to the target word f_i. For instance, $a = (4, 1, 1, 3, 2)$ for the alignment in Figure 2.

4. MODEL 3P

Statistical alignment models for word-to-word alignment are well-studied in SMT literature. However, in *step 3* of our approach outlined in Section 1, aligning words in the source language with phonemes in the target language is required. One method for automatically obtaining such word-to-phoneme alignments is to use word-to-word alignment models from SMT. Authors in [10] used a perfect phoneme transcription on the target side, and ran the word-to-word aligner GIZA++ to align English words to Spanish phonemes. The alignment error rate (AER) of the found alignments was comparable to similar experiments with words instead of phonemes on the target side. Our experiments in Section 6 suggest that significantly better results can be achieved by using our new alignment model *Model 3P* for word-to-phoneme alignment, in particular with respect to word segmentation quality. *Model 3P* extends the IBM Model 3 by additional dependencies for the *translation probabilities* $t(\cdot|\cdot)$ and a set of *word length probabilities* $o(\cdot|\cdot)$. The generative process upon which it is based can be described as follows:

1. For each source word e_i indexed by $i = 1, 2, ..., l$, choose the fertility ϕ_i with probability $n(\phi_i|e_i)$.
2. Choose the number ϕ_0 of "spurious" target words to be generated from $e_0 =$ `NULL`, using probability p_1 and the sum of fertilities from *step 1*.
3. Let $m = \sum_{i=0}^{l} \phi_i$.
4. For each $i = 1, 2, ..., l$, and each $k = 1, 2, ..., \phi_i$, choose a target word position π_{ik} with probability $d(\pi_{ik}|i, l, m)$.

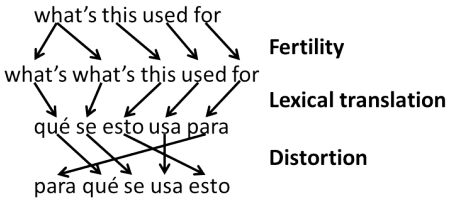


Fig. 3. Generative process in IBM Model 3.

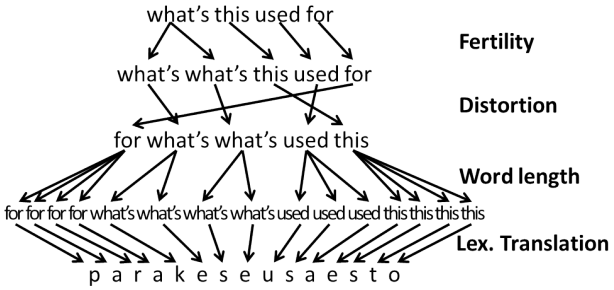
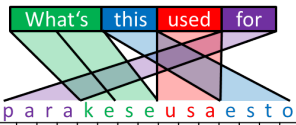


Fig. 4. Generative process in Model 3P.

5. For each $k = 1, 2, ..., \phi_0$, choose a word position π_{0k} from the $\phi_0 - k + 1$ remaining vacant positions in $1, 2, ..., m$, for a total probability of $1/\phi_0!$.
6. For each $i = 0, 1, ..., l$, and each $k = 1, 2, ..., \phi_i$, choose the word length ψ_{ik} with probability $o(\psi_{ik}|e_i)$.
7. For each $i = 0, 1, ..., l$, and each $k = 1, 2, ..., \phi_i$, and each $j = 1, 2, ..., \psi_{ik}$, choose a target phoneme τ_{ikj} with probability $t(\tau_{ikj}|e_i, j)$.
8. Output the target phoneme sequence with phonemes τ_{ikj} in positions π_{ik} $(0 \leq i \leq l, 1 \leq k \leq \phi_i, 1 \leq j \leq \psi_{ik})$.

Besides the fact that *Model 3P* skips *step 4* of IBM Model 3 (lexical translation), both models are identical until



English word position (i)	4	4	4	4	1	1	1	1	3	3	3	2	2	2	2
Target word position (π_{ik})	1	×	×	×	2	×	3	×	4	×	×	5	×	×	×
Target word length (ψ_{ik})	4	×	×	×	2	×	2	×	3	×	×	4	×	×	×
Phoneme position in target word (j)	1	2	3	4	1	2	1	2	1	2	3	1	2	3	4

Fig. 5. Alignments in Model 3P.

applying the distortion model (*step 6* or *step 5*, respectively). At this point, we can regard the target sequence in *Model 3P* as a sequence of anonymous tokens; each is a placeholder for a target word. In *step 6*, we decide for each of these tokens how many phonemes they produce according to the *word length probabilities* $o(\cdot|\cdot)$. The next step fills in the phonemes itself, depending on the source word e_i and their phoneme position j in the target word. Figure 4 illustrates an instance of the generative process of *Model 3P*.

In *Model 3P*, an alignment $A \in \{\mathbb{N} \cup \{\times\}\}^{4\times m}$ is a matrix rather than an integer vector like in IBM Model 3. It captures model decisions made in the fertility and word length step, which would be hidden in an integer vector representation:

- A_{0j}: English word position connected to the j-th target phoneme
- A_{1j}: Position of the target word belonging to the j-th target phoneme
- A_{2j}: Word length in phonemes of the target word A_{1j}
- A_{3j}: Phoneme position of the j-th target phoneme in the corresponding target word

An example alignment is shown in Figure 5. The word boundary between the 6th and 7th phoneme would be not reconstructible in a simple integer vector representation. Equation 2 expresses *Model 3P* as a general formula:

$$\begin{aligned} P(A, f|e) = &\binom{k-\phi_0}{\phi_0} \cdot {p_0}^{k-2\phi_0} \cdot {p_1}^{\phi_0} \\ &\cdot \prod_{i=1}^{l} (n(\phi_i|e_i) \cdot \phi_i!) \cdot \prod_{j=1}^{m} t(f_j|e_{A_{0j}}, A_{3j}) \\ &\cdot \prod_{j:A_{0j}\neq 0, A_{1j}\neq\times}^{m} (d(A_{1j}|A_{0j}, l, k) \cdot o(A_{2j}|e_{A_{0j}})) \end{aligned} \tag{2}$$

5. EXPERIMENTAL SETUP

5.1. Corpus

Our results are based on experiments with a subset of the English-Spanish Basic Travel Expression Corpus (BTEC) [2]. This parallel corpus consists of conversations considered to be useful for people traveling in another country. We removed sentences longer than 50 words or sentence pairs exceeding a word ratio of 9:1 and ended up with 123k sentence pairs and vocabulary sizes of 12k for English and 20k for Spanish. In our experiments, we segment Spanish (unseen target language) phoneme sequences into word units with the help of their English (source language) written translations. For initial experiments, we use correct phoneme sequences, which we generated by replacing the Spanish words with their pronunciations and removing word boundary markers. The pronunciations were taken from a pronunciation dictionary or generated with a Spanish grapheme-to-phoneme model. The Spanish phoneme set consists of 35 phonemes.

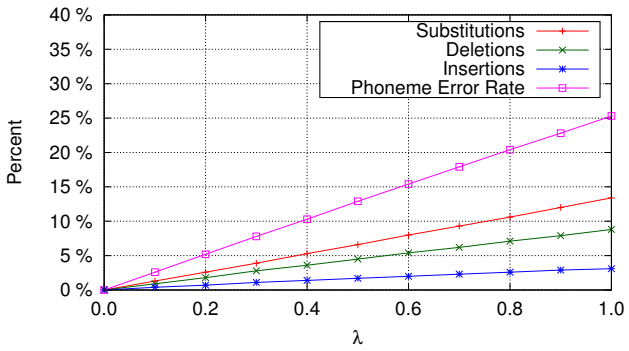


Fig. 6. Phoneme Error Rate over λ.

5.2. Phoneme Errors

We intersperse the correct phoneme sequences with phoneme errors to investigate the performance of different word segmentation approaches depending on the underlying phoneme error rate (PER). In order to imitate recognition errors realistically, we trained a phoneme recognizer on the Spanish *GlobalPhone* corpus [16] and used the NIST sclite scoring and evaluation tool [17] to create its confusion matrix $R \in \mathbb{R}^{36\times 36}$. R_{so} contains the probability $P_R(o|s)$ that the phoneme recognizer confuses the stimulus phoneme s with the observed phoneme o (substitution). An additional row and a column model insertions and deletions, so that all elements in a row sum up to 1 and induce a probability distribution. The Spanish phoneme recognizer has a PER of 25.3%. We smooth R with $\lambda \in [0, 1]$ to control the PER. We obtain a disturbed phoneme sequence by replacing each phoneme s in the perfect phoneme transcription with a phoneme o with the probability $P_\lambda(o|s)$. Figure 6 shows, that the resulting PER is linear in λ. Although errors made by real phoneme recognizers may not be context-independent like in this approach, we feel that it is sufficient to provide a proof of concept for *Model 3P* under conditions with phoneme errors.

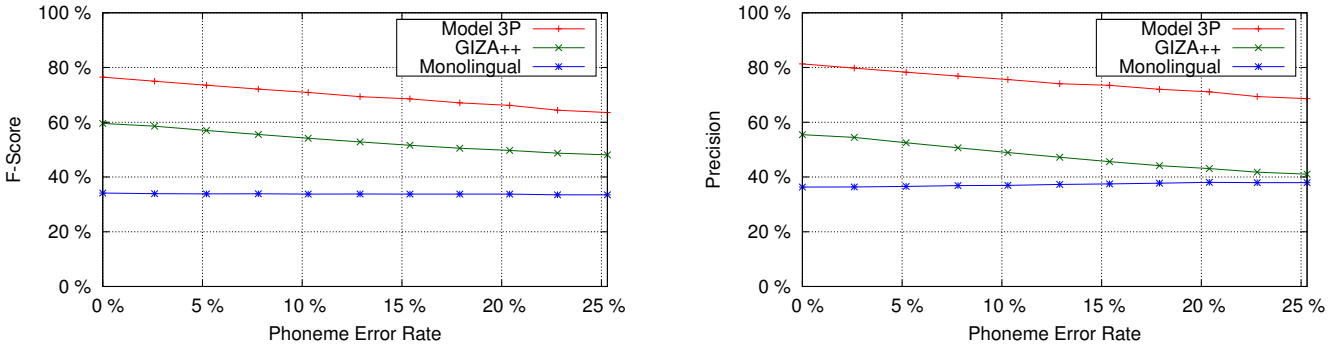
6. EVALUATION

6.1. Systems

We compare the performance of three different unsupervised approaches to word segmentation.

1. *Adaptor Grammars*. We use the implementation from [9] with the *colloc-syllable* grammar representatively for unsupervised monolingual word segmentation methods. This grammar requires a mapping of the phonemes to a vowel and a consonant set.

2. *GIZA++*. The 2nd system uses GIZA++ [11] to obtain word-to-phoneme alignments in a parallel corpus inducing a word segmentation, similar to [10]. Setting the



(a) F-Score (b) Precision

Fig. 7. Word segmentation quality over the Phoneme Error Rate.

GIZA++ parameters *maxfertility* to 12, *deficientdistortionmodelforemptyword* to 1, and *empropforemptyword* to 0.1 has been empirically proven to enhance word-to-phoneme alignment quality.

3. *Model 3P*. The 3rd system is based on our multi-threaded implementation of *Model 3P*[1]. The alignments were found after 10 iterations of the expectation maximization algorithm. The *Model 3P* parameters were initialized using the alignments found by the 2nd system (*GIZA++*), similar to the parameter transfer between the models of the IBM Model hierarchy. The M-step conducts 8,500 iterations of a Genetic Algorithm [18] with the help of the EvA 2 Toolkit [19]. Inserting, removing, and moving word boundaries and realigning a target word to a new source word defined the set of possible mutations. A single-point cross-over operator was used.

6.2. Alignment Performance

Figure 8 shows the alignment performance of both, GIZA++ and *Model 3P* over the PER. The reference alignments were generated by running GIZA++ with default parameters on the word level, and then replacing the Spanish words with their pronunciations afterwards. The Alignment Error Rate (AER) [14] of GIZA++'s word-to-phoneme alignments is up to 13.6% higher than for *Model 3P* alignments. The AER for both systems increase proportionally with the PER, GIZA++ slightly more rapidly than *Model 3P*.

6.3. Word Segmentation Performance

The quality of the found word segmentations is summarized in Figure 7 for all three systems. On correct phoneme sequences, we achieve an F-Score [20] of 76.5%, which is a great improvement over the other systems (34.1% with the monolingual approach, 59.5% with GIZA++). The word accuracy of *Model 3P* is 90.1%. Again, we observe a linear decline for both cross-lingual approaches with increasing PER. Given the definition of vowel and consonant sets, the monolingual approach seems to be more robust against recognition errors, but comes with not more than 34.1% F-Score. With a PER of 25.3%, word segmentations based on the monolingual approach and GIZA++ have approximately the same precision. Applying *Model 3P* outperforms both other methods regardless of the PER. The accuracy is still 83.9% on a phoneme sequence containing the 25.3% errors produced by our phoneme recognizer. From the 123k sentence, each 5th sentence contains completely correct segmented word units with *Model 3P*. This means that found word units often correspond to real Spanish words. With GIZA++, it is only each 12th. A look at some segmentation outputs of *Model 3P* reveals that it has rather problems in segmenting short Spanish words containing one or two phonemes such as "la" and "a" than longer ones. Furthermore, off-by-one errors appear at morphological boundaries and at words with only one phoneme. The segmentation of long and rare words is successful if adjacent words are frequent.

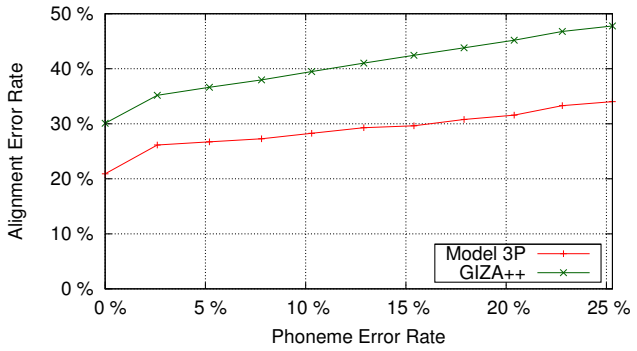


Fig. 8. Alignment Error Rates over the underlying Phoneme Error Rate.

[1] Available at `http://pisa.googlecode.com/`

7. CONCLUSION AND FUTURE WORK

The word segmentation problem describes the task of segmenting phoneme sequences into word units. We have investigated three different unsupervised algorithms for automatically finding word boundaries in phonetic transcriptions. We showed that using information from another language rather than a pure monolingual approach helps to find better segmentations on correct phoneme sequences. A simple way to incorporate cross-lingual information is to apply word-to-word alignment models from SMT to align words of the other language to the phonemes of the target language. However, with these word-to-word alignment models the word segmentation precision is not significantly higher than in the monolingual approach when phoneme recognition errors are common. Therefore we proposed the new alignment model *Model 3P* for cross-lingual word-to-phoneme alignment, which extends the generative process of IBM Model 3 by a word length step and additional dependencies for the lexical translation probabilities. With this new model, we obtain considerably better word segmentations than with both previous methods. Using *Model 3P* for the alignment between English words and Spanish phonemes outperformed a state-of-the-art monolingual word segmentation approach on the BTEC corpus by up to 42% absolute in F-Score on the phoneme level and a GIZA++ alignment based on IBM Model 3 by up to 17%. We report a word accuracy of 90.1% on correct phoneme transcriptions, and still 83.9% on a phoneme sequence containing the 25.3% errors produced by our phoneme recognizer.

In the near future, we plan to explore our approach on other language pairs. Furthermore, we are working on an algorithm that iteratively extracts phoneme sequences, induces phonetic transcriptions of words and compensates for alignment and phoneme recognition errors. According to the *steps 4a)* and *4b)* of the scenario in Section 1, our goal is to use the resulting word sequences in the training process of an MT system and bootstrap pronunciation dictionaries for under-resourced languages, and those that are not written at all.

8. REFERENCES

[1] M. Johnson and S. Goldwater, "Improving Nonparameteric Bayesian Inference: Experiments on Unsupervised Word Segmentation with Adaptor Grammars," in *ACL-HLT*, 2009, pp. 317–325.

[2] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating Corpora for Speech-to-Speech Translation," in *Eurospeech*, 2003.

[3] R. G. Gordon and B. F. Grimes, *Ethnologue: Languages of the World*, SIL International, 15th edition, 2005.

[4] T. Schultz and K. Kirchhoff, Eds., *Multilingual Speech Processing*, Academic Press, Amsterdam, 2006.

[5] D. Nettle and S. Romaine, *Vanishing Voices: The Extinction of the World's Languages*, Oxford University Press, 2000.

[6] L. Besacier, B. Zhou, and Y. Gao, "Towards Speech Translation of Non-Written Languages," in *SLT*, 2006, pp. 222–225.

[7] C.Y. Kit, *Unsupervised Lexical Learning as Inductive Inference*, Ph.D. thesis, University of Sheffield, 2000.

[8] J. Goldsmith, "An Algorithm for the Unsupervised Learning of Morphology," *Natural Language Engineering*, vol. 12, pp. 353–371, 2006.

[9] M. Johnson, "Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure," in *ACL-HLT*, 2008, pp. 398–406.

[10] S. Stüker and A. Waibel, "Towards Human Translations Guided Language Discovery for ASR Systems," in *SLTU*, 2008.

[11] F. J. Och and H. Ney, "Improved Statistical Alignment Models," in *ACL*, 2000, pp. 440–447.

[12] S. Vogel, H. Ney, and C. Tillmann, "HMM-based Word Alignment in Statistical Translation," in *COLING*, 1996, pp. 836–841.

[13] P. Brown, S. Della Pietra, V. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.

[14] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, pp. 19–51, 2003.

[15] K. Knight, "A Statistical MT Tutorial Workbook," in *JHU Summer Workshop*, 1999.

[16] T. Schultz, "GlobalPhone: A Multilingual Speech and Text Database Developed at Karlsruhe University," in *ICSLP*, 2002.

[17] J. Fiscus, "Speech Recognition Scoring Toolkit ver. 2.3 (sctk)," 2007.

[18] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[19] M. Kronfeld, H. Planatscher, and A. Zell, "The EvA2 Optimization Framework," *Learning and Intelligent Optimization*, pp. 247–250, 2010.

[20] C. J. Van Rijsbergen, *Information Retrieval*, Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.