Modeling with Structures in Statistical Machine Translation

Ye-Yi Wang and **Alex Waibel**
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{yyw,waibel}@cs.cmu.edu

Abstract

Most statistical machine translation systems employ a word-based alignment model. In this paper we demonstrate that word-based alignment is a major cause of translation errors. We propose a new alignment model based on shallow phrase structures, and the structures can be automatically acquired from parallel corpus. This new model achieved over 10% error reduction for our spoken language translation task.

1 Introduction

Most (if not all) statistical machine translation systems employ a word-based alignment model (Brown et al., 1993; Vogel, Ney, and Tillman, 1996; Wang and Waibel, 1997), which treats words in a sentence as independent entities and ignores the structural relationship among them. While this independence assumption works well in speech recognition, it poses a major problem in our experiments with spoken language translation between a language pair with very different word orders. In this paper we propose a translation model that employs shallow phrase structures. It has the following advantages over word-based alignment:

- Since the translation model can directly depict phrase reordering in translation, it is more accurate for translation between languages with different word (phrase) orders.
- The decoder of the translation system can use the phrase information and extend hypothesis by phrases (multiple words), therefore it can speed up decoding.

The paper is organized as follows. In section 2, the problems of word-based alignment models are discussed. To alienate these problems, a new alignment model based on shallow phrase structures is introduced in section 3. In section 4, a grammar inference algorithm is presented that can automatically acquire the phrase structures used in the new model. Translation performance is then evaluated in section 5, and conclusions are presented in section 6.

2 Word-based Alignment Model

In a word-based alignment translation model, the transformation from a sentence at the source end of a communication channel to a sentence at the target end can be described with the following random process:

1. Pick a length for the sentence at the target end.
2. For each word position in the target sentence, align it with a source word.
3. Produce a word at each target word position according to the source word with which the target word position has been aligned.

IBM Alignment Model 2 is a typical example of word-based alignment. Assuming a sentence $\mathbf{s} = s_1, \cdots, s_l$ at the source of a channel, the model picks a length m of the target sentence $\mathbf{t}$ according to the distribution $P(m \mid \mathbf{s}) = \epsilon$, where ϵ is a small, fixed number. Then for each position i $(0 < i \leq m)$ in $\mathbf{t}$, it finds its corresponding position a_i in $\mathbf{s}$ according to an *alignment* distribution $P(a_i \mid i, a_1^{i-1}, m, \mathbf{s}) = a(a_i \mid i, m, l)$. Finally, it generates a word t_i at the position i of $\mathbf{t}$ from the source word s_{a_i} at the aligned position a_i, according to a *translation* distribution $P(t_i \mid t_1^{i-1}, a_1^m, \mathbf{s}) = t(t_i \mid s_{a_i})$.

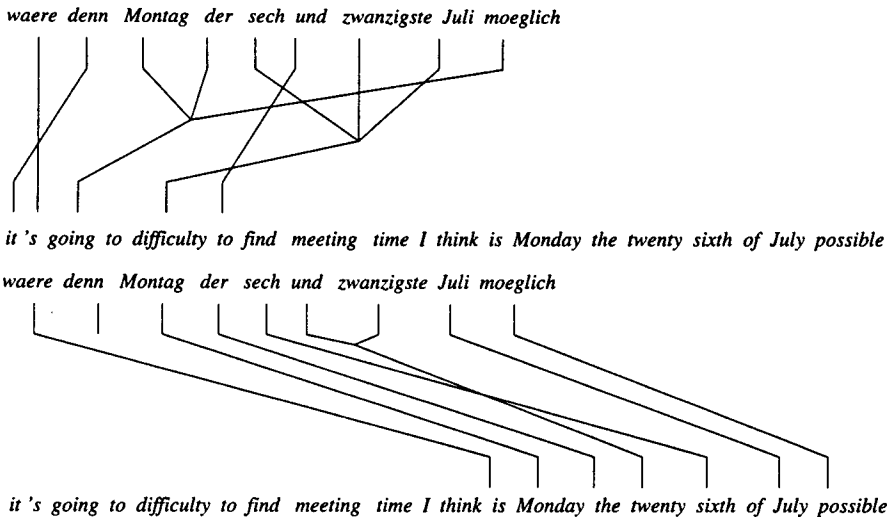

Figure 1: Word Alignment with deletion in translation: the top alignment is the one made by IBM Alignment Model 2, the bottom one is the 'ideal' alignment.

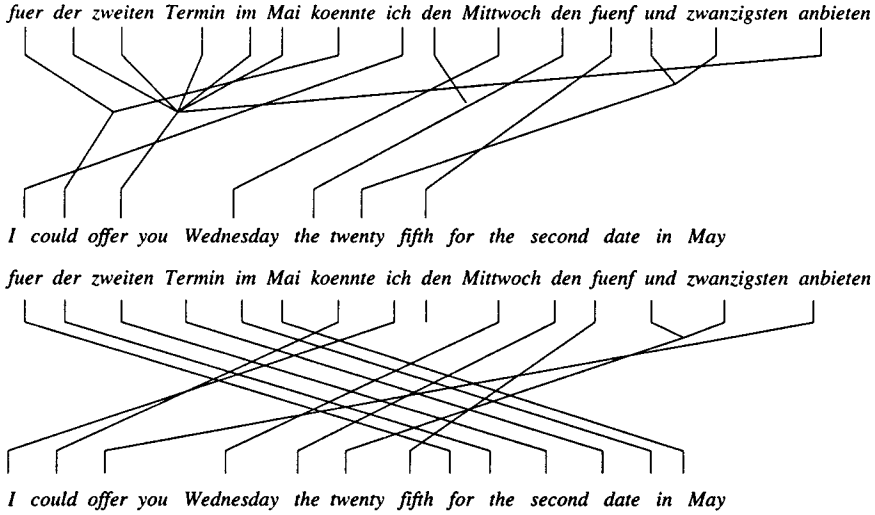

Figure 2: Word Alignment of translation with different phrase order: the top alignment is the one made by IBM Alignment Model 2, the bottom one is the 'ideal' alignment.

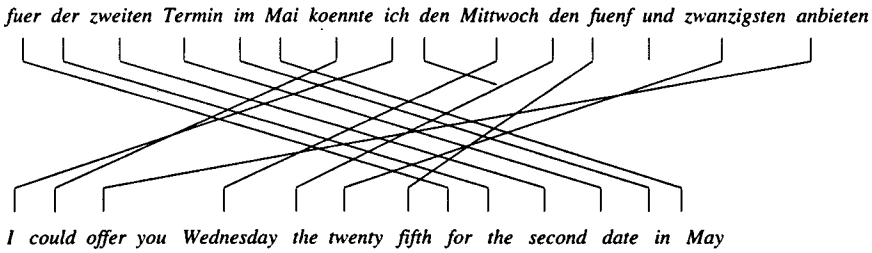

Figure 3: Word Alignment with Model 1 for one of the previous examples. Because no alignment probability penalizes the long distance phrase reordering, it is much closer to the 'ideal' alignment.

Therefore, $P(\mathbf{t} \mid \mathbf{s})$ is the sum of the probabilities of generating $\mathbf{t}$ from $\mathbf{s}$ over all possible alignments A, in which the position i in $\mathbf{t}$ is aligned with the position a_i in $\mathbf{s}$:

$$\begin{aligned} & P(\mathbf{t} \mid \mathbf{s}) \\ = & \epsilon \sum_{a_1=0}^{l} \cdots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(t_j \mid s_{a_j}) a(a_j \mid j, l, m) \\ = & \epsilon \prod_{j=1}^{m} \sum_{i=0}^{l} t(t_j \mid s_i) a(i \mid j, l, m) \end{aligned} \quad (1)$$

A word-based model may have severe problems when there are deletions in translation (this may be a result of erroneous sentence alignment) or the two languages have different word orders, like English and German. Figure 1 and Figure 2 show some problematic alignments between English/German sentences made by IBM Model 2, together with the 'ideal' alignments for the sentences. Here the alignment parameters penalize the alignment of English words with their German translation equivalents because the translation equivalents are far away from the words.

An experiment reveals how often this kind of "skewed" alignment happens in our English/German scheduling conversation parallel corpus (Wang and Waibel, 1997). The experiment was based on the following observation: IBM translation Model 1 (where the alignment distribution is uniform) and Model 2 found similar Viterbi alignments when there were no movements or deletions, and they predicted very different Viterbi alignments when the skewness was severe in a sentence pair, since the alignment parameters in Model 2 penalize the long distance alignment. Figure 3 shows the Viterbi alignment discovered by Model 1 for the same sentences in Figure 2[1].

We measured the distance of a Model 1 alignment $\mathbf{a^1}$ and a Model 2 alignment $\mathbf{a^2}$ as $\sum_{i=1}^{|\mathbf{g}|} |a_i^1 - a_i^2|$. To estimate the skewness of the corpus, we collected the statistics about the percentage of sentence pairs (with at least five words in a sentence) with Model 1 and Model 2 alignment distance greater than $1/4, 2/4, 3/4, \cdots, 10/4$ of the target sentence length. By checking the Viterbi alignments made by both models, it is almost certain that whenever the distance is greater that 3/4 of the target sentence length, there is either a movement or a deletion in the sentence pair. Figure 4 plots this statistic — around 30% of the sentence pairs in our training data have some degree of skewness in alignments.

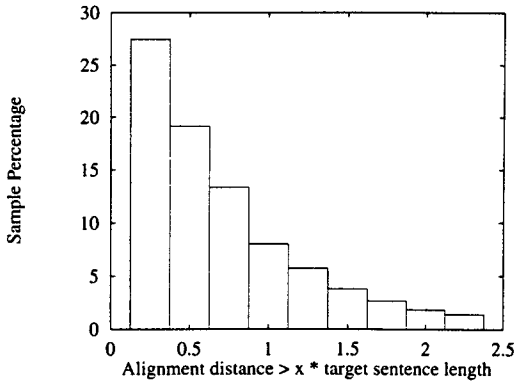

Figure 4: Skewness of Translations

3 Structure-based Alignment Model

To solve the problems with the word-based alignment models, we present a structure-based alignment model here. The idea is to directly model the phrase movement with a rough alignment, and then model the word alignment within phrases with a detailed alignment.

Given an English sentence $\mathbf{e} = e_1 e_2 \cdots e_l$, its German translation $\mathbf{g} = g_1 g_2 \cdots g_m$ can be generated by the following process:

1. Parse $\mathbf{e}$ into a sequence of phrases, so

$$\begin{aligned} E &= (e_{11}, e_{12}, \cdots, e_{1l_1})(e_{21}, e_{22}, \cdots, e_{2l_2}) \cdots \\ &\quad (e_{n1}, e_{n2}, \cdots, e_{nl_n}) \\ &= E_0 E_1 E_2 \cdots E_n, \end{aligned}$$

   where E_0 is a null phrase.

2. With the probability $P(q \mid \mathbf{e}, E)$, determine $q \leq n+1$, the number of phrases in $\mathbf{g}$. Let $G_1 \cdots G_q$ denote these q phrases. Each source phrase can be aligned with at most one target phrase. Unlike English phrases, words in a German phrase do not

[1]The better alignment on a given pair of sentences does not mean Model 1 is a better model. Non-uniform alignment distribution is desirable. Otherwise, language model would be the only factor that determines the source sentence word order in decoding.

have to form a consecutive sequence. So $\mathbf{g}$ may be expressed with something like $\mathbf{g} = g_{11}g_{12}g_{21}g_{13}g_{22}\cdots$, where g_{ij} represents the j-th word in the i-th phrase.

3. For each German phrase $G_i, 0 \leq i < q$, with the probability $P(r_i \mid i, r_0^{i-1}, E, \mathbf{e})$, align it with an English phrase E_{r_i}.
4. For each German phrase $G_i, 0 \leq i < q$, determine its beginning position b_i in $\mathbf{g}$ with the distribution $P(b_i \mid i, b_0^{i-1}, r_0^q, \mathbf{e}, E)$.
5. Now it is time to generate the individual words in the German phrases through *detailed alignment*. It works like IBM Model 4. For each word e_{ij} in the phrase E_i, its fertility ϕ_{ij} has the distribution $P(\phi_{ij} \mid i, j, \phi_{i1}^{j-1}, \phi_0^{i-1}, b_0^q, r_0^q, e, E)$.
6. For each word e_{ij} in the phrase E_i, it generates a tablet $\tau_{ij} = \{\tau_{ij1}, \tau_{ij2}, \cdots \tau_{ij\phi_{ij}}\}$ by generating each of the words in τ_{ij} in turn with the probability $P(\tau_{ijk} \mid \tau_{ij1}^{k-1}, \tau_{i1}^{j-1}, \tau_0^{i-1}, \phi_0^l, b_0^q, r_0^q, e, E)$ for the k-th word in the tablet.
7. For each element τ_{ijk} in the tablet τ_{ij}, the permutation π_{ijk} determines its position in the target sentence according to the distribution $P(\pi_{ijk} \mid \pi_{ij1}^{k-1}, \pi_{i1}^{j-1}, \pi_0^{i-1}, \tau_0^l, \phi_0^l, b_0^q, r_0^q, e, E)$.

We made the following independence assumptions:

1. The number of target sentence phrases depends only on the number of phrases in the source sentence:
   $P(q \mid \mathbf{e}, E) = p_n(q \mid n)$
2. $P(r_i \mid i, r_0^{i-1}, E, \mathbf{e})$
   $= a(r_i \mid i) \times \prod_{0 \leq j < i}(1 - \delta(r_i, r_j))$
   where $\delta(x, y) = 1$ when $x = y$, and $\delta(x, y) = 0$ otherwise.
   This assumption states that $P(r_i \mid i, r_0^{i-1}, E, \mathbf{e})$ depends on i and r_i. It also depends on r_0^{i-1} with the factor $\prod_{0<j<i}(1-\delta(r_i, r_j))$ to ensure that each English phrase is aligned with at most one German phrase.
3. The beginning position of a target phrase depends on its distance from the beginning position of its preceding phrase, as well as the length of the source phrase aligned with the preceding phrase:
   $P(b_i \mid i, b_0^{i-1}, r_0^q, \mathbf{e}, E)$
   $= \alpha(b_i - b_{i-1} \mid |E_{r_{i-1}}|) = \alpha(\Delta_i \mid |E_{r_{i-1}}|)$
4. The fertility and translation tablet of a source word depend on the word only:
   $P(\phi_{ij} \mid i, j, \phi_{i1}^{j-1}, \phi_0^{i-1}, b_0^q, r_0^q, e, E)$
   $= n(\phi_{ij} \mid e_{ij})$
   $P(\tau_{ijk} \mid \tau_{ij1}^{k-1}, \tau_{i1}^{j-1}, \tau_0^{i-1}, \phi_0^l, b_0^q, r_0^q, e, E)$
   $= t(\tau_{ijk} \mid e_{ij})$
5. The leftmost position of the translations of a source word depends on its distance from the beginning of the target phrase aligned with the source phrase that contains that source word. It also depends on the identity of the phrase, and the position of the source word in the source phrase.
   $P(\pi_{ij1} \mid \pi_{i1}^{j-1}, \pi_0^{i-1}, \tau_0^l, \phi_0^l, b_0^q, r_0^q, e, E)$
   $= d_1(\pi_{ij1} - b_i \mid E_i, j)$

   For a target word τ_{ijk} other than the leftmost τ_{ij1} in the translation tablet of the source e_{ij}, its position depends on its distance from the position of another tablet word $\tau_{ij(k-1)}$ closest to its left, the class of the target word τ_{ijk}, and the fertility of the source word e_{ij}.
   $P(\pi_{ijk} \mid \pi_{ij1}^{k-1}, \pi_{i1}^{j-1}, \pi_0^{i-1}, \tau_0^l, \phi_0^l, b_0^q, r_0^q, e, E)$
   $= d_2(\pi_{ijk} - \pi_{ij(k-1)} \mid \mathcal{G}(\tau_{ijk}), \phi_{ij})$
   here $\mathcal{G}(g)$ is the equivalent class for g.

3.1 Parameter Estimation

EM algorithm was used to estimate the seven types of parameters: p_n, a, α, ϕ, τ, d_1 and d_2. We used a subset of probable alignments in the EM learning, since the total number of alignments is exponential to the target sentence length. The subset was the neighboring alignments (Brown et al., 1993) of the Viterbi alignments discovered by Model 1 and Model 2. We chose to include the Model 1 Viterbi alignment here because the Model 1 alignment is closer to the "ideal" when strong skewness exists in a sentence pair.

4 Finding the Structures

It is of little interest for the structure-based alignment model if we have to manually find

the language structures and write a grammar for them, since the primary merit of statistical machine translation is to reduce human labor. In this section we introduce a grammar inference technique that finds the phrases used in the structure-based alignment model. It is based on the work in (Ries, Buø, and Wang, 1995), where the following two operators are used:

1. **Clustering:** Clustering words/phrases with similar meanings/grammatical functions into equivalent classes. The mutual information clustering algorithm(Brown et al., 1992) were used for this.

2. **Phrasing:** The equivalent class sequence $c_1, c_2, \cdots c_k$ forms a phrase if

$$P(c_1, c_2, \cdots c_k) \log \frac{P(c_1, c_2, \cdots c_k)}{P(c_1)P(c_2)\cdots P(c_k)} > \theta,$$

where θ is a threshold. By changing the threshold, we obtain a different number of phrases.

The two operators are iteratively applied to the training corpus in alternative steps. This results in hierarchical phrases in the form of sequences of equivalent classes of words/phrases.

Since the algorithm only uses a monolingual corpus, it often introduces some language-specific structures resulting from biased usages of a specific language. In machine translation we are more interested in cross-linguistic structures, similar to the case of using interlingua to represent cross-linguistic information in knowledge-based MT.

To obtain structures that are common in both languages, a bilingual mutual information clustering algorithm (Wang, Lafferty, and Waibel, 1996) was used as the clustering operator. It takes constraints from parallel corpus. We also introduced an additional constraint in clustering, which requires that words in the same class must have at least one common potential part-of-speech.

Bilingual constraints are also imposed on the phrasing operator. We used bilingual heuristics to filter out the sequences acquired by the phrasing operator that may not be common in multiple languages. The heuristics include:

1. **Average Translation Span:** Given a phrase candidate, its average translation span is the distance between the leftmost and the rightmost target positions aligned with the words inside the candidate, averaged over all Model 1 Viterbi alignments of sample sentences. A candidate is filtered out if its average translation span is greater than the length of the candidate multiplied by a threshold. This criterion states that the words in the translation of a phrase have to be close enough to form a phrase in another language.

2. **Ambiguity Reduction:** A word occurring in a phrase should be less ambiguous than in other random context. Therefore a phrase should reduce the ambiguity (uncertainty) of the words inside it. For each source language word class c, its translation entropy is defined as $\sum_g t(g \mid c) \log(g \mid c)$. The average per source class entropy reduction induced by the introduction of a phrase P is therefore

$$\frac{1}{|P|} \sum_{c \in P} [\sum_g t(g \mid c) \log t(g \mid c) - \sum_g t(g \mid c, P) \log t(g \mid c, P)]$$

A threshold was set up for minimum entropy reduction.

By applying the clustering operator followed with the phrasing operator, we obtained shallow phrase structures partly shown in Figure 5.

Given a set of phrases, we can deterministically parse a sentence into a sequence of phrases by replacing the leftmost unparsed substring with the longest matching phrase in the set.

5 Evaluation and Discussion

We used the Janus English/German scheduling corpus (Suhm et al., 1995) to train our phrase-based alignment model. Around 30,000 parallel sentences (400,000 words altogether for both languages) were used for training. The same data were used to train Simplified Model 2 (Wang and Waibel, 1997) and IBM Model 3 for performance comparison. A larger English monolingual corpus with around 0.5 million words was used for the training of a bigram

```
[Sunday Monday...] [afternoon morning...]
[Sunday Monday...] [at by...] [one two...]
[Sunday Monday...] [the every each...] [first second third...]
[Sunday Monday...] [the every each...] [twenty depending remaining]
[Sunday Monday...] [the every each...] [eleventh thirteenth...]
[Sunday Monday...] [in within...] [January February...]
[January February...] [first second third...] [at by...]
[January February...] [first second third...]
[January February...] [the every each...] [first second third...]
[I he she itself] [have propose remember hate...]
[eleventh thirteenth...] [after before around] [one two three...]
```

Figure 5: Example of Acquired Phrases. Words in a bracket form a cluster, phrases are cluster sequences. Ellipses indicate that a cluster has more words than those shown here.

Model	Correct	OK	Incorrect	Accuracy
Model 2	284	87	176	59.9%
Model 3	98	45	57	60.3%
S. Model	303	96	148	64.2%

Table 1: Translation Accuracy: a correct translation gets one credit, an okay translation gets 1/2 credit, an incorrect one gets 0 credit. Since the IBM Model 3 decoder is too slow, its performance was not measured on the entire test set.

language model. A preprocessor splited German compound nouns. Words that occurred only once were taken as unknown words. This resulted in a lexicon of 1372 English and 2202 German words. The English/German lexicons were classified into 250 classes in each language and 560 English phrases were constructed upon these classes with the grammar inference algorithm described earlier.

We limited the maximum sentence length to be 20 words/15 phrases long, the maximum fertility for non-null words to be 3.

5.1 Translation Accuracy

Table 1 shows the end-to-end translation performance. The structure-based model achieved an error reduction of around 12.5% over the word-based alignment models.

5.2 Word Order and Phrase Alignment

Table 2 shows the alignment distribution for the first German word/phrase in Simplified Model 2 and the structure-based model. The probability mass is more scattered in the structure-based model, reflecting the fact that English and German have different phrase orders. On the other hand, the word based model tends to align a target word with the source words at similar positions, which resulted in many incorrect alignments, hence made the word translation probability t distributed over many unrelated target words, as to be shown in the next subsection.

5.3 Model Complexity

The structure-based model has 3,081,617 free parameters, an increase of about 2% over the 3,022,373 free parameters of Simplified Model 2. This small increase does not cause over-fitting, as the performance on the test data suggests. On the other hand, the structure-based model is more accurate. This can be illustrated with an example of the translation probability distribution of the English word "I". Table 3 shows the possible translations of "I" with probability greater than 0.01. It is clear that the structure-based model "focuses" better on the correct translations. It is interesting to note that the German translations in Simplified Model 2 often appear at the beginning of a sentence, the position where "I" often appears in English sentences. It is the biased word-based alignments that pull the unrelated words together and increase the translation uncertainty.

We define the *average translation entropy* as

$$\sum_{i=0}^{m} P(e_i) \sum_{j=1}^{n} -t(g_j \mid e_i) \log t(g_j \mid e_i).$$

j	0	1	2	3	4	5	6	7	8	9	...
$a_{M2}(j \mid 1)$	0.04	0.86	0.054	0.025	0.008	0.005	0.004	0.002	3.3×10^{-4}	2.9×10^{-4}	...
$a_{SM}(j \mid 1)$	0.003	0.29	0.25	0.15	0.07	0.11	0.05	0.04	0.02	0.01	...

Table 2: The alignment distribution for the first German word/phrase in Simplified Model 2 and in the structure-based model. The second distribution reflects the higher possibility of phrase reordering in translation.

$t_{M2}(* \mid \text{I})$		$t_{SM}(* \mid \text{I})$	
ich	0.708	ich	0.988
da	0.104	mich	0.010
am	0.024		
das	0.022		
dann	0.022		
also	0.019		
es	0.011		

Table 3: The translation distribution of "I". It is more uncertain in the word-based alignment model because the biased alignment distribution forced the associations between unrelated English/German words.

(m, n are English and German lexicon size.) It is a direct measurement of word translation uncertainty. The average translation entropy is 3.01 bits per source word in Simplified Model 2, 2.68 in Model 3, and 2.50 in the structured-based model. Therefore information-theoretically the complexity of the word-based alignment models is higher than that of the structure-based model.

6 Conclusions

The structure-based alignment directly models the word order difference between English and German, makes the word translation distribution focus on the correct ones, hence improves translation performance.

7 Acknowledgements

We would like to thank the anonymous COLING/ACL reviewers for valuable comments. This research was partly supported by ATR and the Verbmobil Project. The views and conclusions in this document are those of the authors.

References

Brown, P. F., S. A. Della-Pietra, V. J Della-Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Brown, P. F., V. J. Della-Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer. 1992. Class-Based N-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Ries, Klaus, Finn Dag Buø, and Ye-Yi Wang. 1995. Improved Language Modelling by Unsupervised Acquisition of Structure. In *ICASSP '95*. IEEE. corrected version available via `http://www.cs.cmu.edu/fies/icassp_95.html`.

Suhm, B., P.Geutner, T. Kemp, A. Lavie, L. Mayfield, A. McNair, I. Rogina, T. Schultz, T. Sloboda, W. Ward, M. Woszczyna, and A. Waibel. 1995. JANUS: Towards multilingual spoken language translation. In *Proceedings of the ARPA Speech Spoken Language Technology Workshop, Austin, TX, 1995.*

Vogel, S., H. Ney, and C. Tillman. 1996. HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the Seventeenth International Conference on Computational Linguistics: COLING-96*, pages 836–841, Copenhagen, Denmark.

Wang, Y., J. Lafferty, and A. Waibel. 1996. Word Clustering with Parallel Spoken Language Corpora. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP'96)*, Philadelphia, USA.

Wang, Y. and A. Waibel. 1997. Decoding Algorithm in Statistical Machine Translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL'97)*, pages 366–372, Madrid, Spain.