Using Senses in HMM Word Alignment

Douwe Gelling and **Trevor Cohn**
Department of Computer Science
University of Sheffield, UK
{d.gelling,t.cohn}@sheffield.ac.uk

Abstract

Some of the most used models for statistical word alignment are the IBM models. Although these models generate acceptable alignments, they do not exploit the rich information found in lexical resources, and as such have no reasonable means to choose better translations for specific senses.

We try to address this issue by extending the IBM HMM model with an extra hidden layer which represents the senses a word can take, allowing similar words to share similar output distributions. We test a preliminary version of this model on English-French data. We compare different ways of generating senses and assess the quality of the alignments relative to the IBM HMM model, as well as the generated sense probabilities, in order to gauge the usefulness in Word Sense Disambiguation.

1 Introduction

Modern machine translation is dominated by statistical methods, most of which are trained on word-aligned parallel corpora (Koehn et al., 2007; Koehn, 2004), which need to be generated separately. One of the most commonly used methods to generate these word alignments is to use the IBM models 1-5, which generate one-directional alignments.

Although the IBM models perform well, they fail to take into account certain situations. For example, if an alignment between two words f_1 and e_1 is considered, and f_1 is an uncommon translation for e_1, the translation probability will be low. It might happen, that an alignment to a different nearby word is preferred by the model. Consider for example the situation where f_1 is 'taal' (Dutch, meaning language), and e_1 is 'tongue'. The translation probability for this may be low, as 'tongue' usually translates as 'tong', meaning the body part. In this case the preference of the alignment model may dominate, leading to the wrong alignment.

Moreover, the standard tools for word alignment fail to make use of the lexical resources that already exist, and which could contribute useful information for the task. In particular, the ontology defined in WordNet (Miller, 1995) could be put to good use. Intuitively, the translation of a word should depend on the sense of the word being used. The current work seeks to explore this idea, by explicitly modeling the senses in the translation process. It does so, by modifying the HMM alignment model to include synsets as an intermediate stage of translation. This would facilitate sharing of translation distributions between words with similar senses that should generate the correct sense. In terms of the example above, one of the senses for 'tongue' will share the translation distribution with 'language', for which we will have more relevant translation probabilities.

As well as performing word alignment this model can be used to generate sense annotations on one side of a parallel corpus, given an alignment, or even generate sense annotations while aligning a corpus. Thus, the model could learn to align a corpus and do WSD at the same time. In this paper, the effect the usage of senses has on alignment is investigated, and the potential usefulness of the model for WSD is explored. In the next section related work is discussed, after which in section 3 the current model is

NAACL-HLT Workshop on the Induction of Linguistic Structure, pages 39–46,
Montréal, Canada, June 3-8, 2012.

discussed.

In section 4 the evaluation of the model is discussed, in two parts. In the first part, the model is evaluated for English-French on gold standard manually aligned data and compared to the results of the base HMM model. In the second part, the model is qualitatively evaluated by inspecting the senses and associated output distributions of selected words.

2 Previous Work

Although most researchers agree that Word Sense Disambiguation (WSD) is a useful field, it hasn't been shown to consistently help in related tasks. Machine Translation is no exception, and whether or not WSD systems can improve performance of MT systems is debated. Furthermore, it is unclear how parallel corpuses can be exploited for WSD systems. In this section we will present a brief overview of related work.

(Carpuat and Wu, 2007) report an improvement in translation quality by incorporating a WSD system directly in a phrase-based translation system. This is in response to earlier work done, where incorporating the output of a traditional WSD system gave disappointing results (Carpuat and Wu, 2005). The WSD task is redefined, to be similar to choosing the correct phrasal translation for a word, instead of choosing a sense from a sense inventory. This system is trained on the same data as the SMT system is.

The output of this model is incorporated into the machine translation system by providing the WSD probabilities for a phrase translation as extra features in a log-linear model (Carpuat and Wu, 2007). This system consistently outperforms the baseline system (the same system, but without WSD component), on multiple metrics, which seems to indicate that WSD can make a useful contribution to machine translation. However, the way the system is set up, it could also be viewed as a way of incorporating translation probabilities of other systems into the phrase-based translation model.

(Chan and Ng, 2007) introduce a system very similar to that of (Carpuat and Wu, 2007), but as applied to hierarchical phrase-based translation. They demonstrate modest improvements in BLEU score over the unmodified system, as well as some qualitative improvements in the output. Here again, the argument could be made that what is being done is not strictly word sense disambiguation, but augmenting the translation system with extra features for some of the phrase translations.

In (Tufiş et al., 2004) parallel corpora and aligned WordNets are exploited for WSD. This is done, by word aligning the parallel texts, and then for every aligned pair, generating a set of wordnet sense codes (ILI codes, or interlingual index codes) for either word, corresponding to the possible senses that word can take. As the wordnets for both languages are linked, if the ILI code of a sense is the same, the sense should be sufficiently similar. Thus, the intersection of both sets of ILI is taken to find an ILI code that is common to both pairs. If such a code is found, it represents the sense index of both words. Otherwise, the closest ILI code to the two most similar ILI codes is found, and that is taken as the sense for the word. The current work however only uses a lexical resource for one of the languages, and as such has fewer places to fail, and less demanding requirements.

Other similar work includes that in (Ng et al., 2003), where a sense-annotated corpus was automatically generated from a parallel corpus. This is done by word-aligning the parallel corpus, and then finding the senses according to WordNet given a list of nouns. Two senses are lumped together if they are translated into the same chinese word. The selection of correct translations is done manually. Only those occurrences of the chosen nouns that translate to one of the chosen chinese words are considered sense-tagged by the translation.

Although similar in approach to what the current system would do, this system uses a much more simple approach to generate sense annotations and it depends on a previously word-aligned corpus, whereas the current approach would integrate alignment and sense-tagging, whis may give a higher accuracy.

3 Senses Model

The current model is based on the HMM alignment model (Vogel et al., 1996), as it is a less complex model than IBM models 3 and above, but still finds acceptable alignments. The HMM alignment model is defined as a HMM model, where the observed

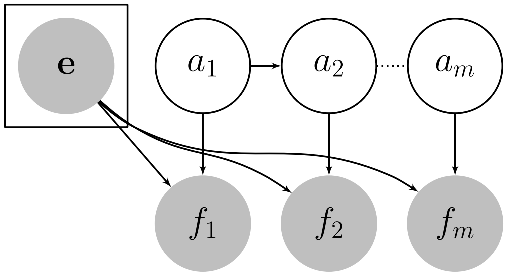

Figure 1: Diagram of HMM model. Arrows indicate dependencies, grey nodes indicate known values, white nodes indicate hidden variables.

variables are the words of a sentence in the French language **f**, and the hidden variables are alignments to words in the English sentence **e**, or to a null state. See figure 1 for a diagram of the standard HMM model. Under this model, French words can align to at most 1 English word. The transition probability is not dependent on the english words themselves, but on the size of jumps between alignments and the length of the English sentence. The probability of the French sentence given the English sentence is:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \prod_{j=1}^{J} p(f_j|e_{a_j}) p(a_j|a_{j-1}, I) \quad (1)$$

Here, **f** and **e** denote the French and English sentences, which have lengths J and I respectively, and **a** denotes an alignment of these two sentences. So, the states in the HMM assign a number from the range $[0, I]$ to each of the positions j in the French sentence, effectively assigning one English word e_{a_j} to each French word f_j, or a NULL translation e_0. The term $p(f_j|e_{a_j})$ is the translation probability of a pair of words, and $p(a_j|a_{j-1}, I)$ gives the transition probability in the HMM.

Here, i is the current state of the HMM, and i' is the previous state of the HMM, each being an index into the English sentence and $p(a_j|a_{j-1}, I)$ is defined as the probability of the gap between i and i'. So, if in an alignment French word 2 is aligned to the 3rd English word, and the next French Word (3) is aligned to the 5th English word, $p(a_j|a_{j-1}, I)$ isn't modelled directly as $p(5|3, I)$, but as $p(5 - 3|I)$.

To implement a dependency on senses in the model an extra hidden layer is added to the HMM model, representing the senses. The probability of a french word then depends on the generated sense, the probability of which depends on the English. The possible senses for a given English word is constrained by an external source, such as WordNet.

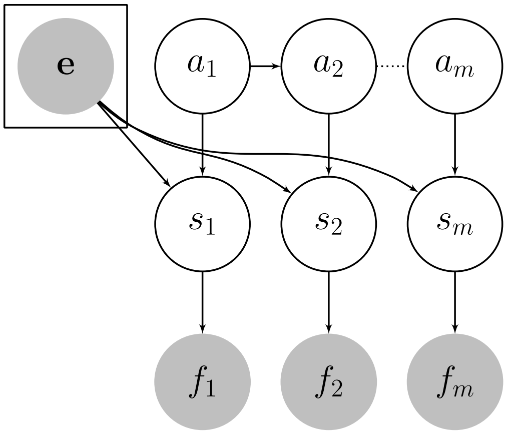

Figure 2: Diagram of SHMM model, with senses generated by the English words. Arrows indicate dependencies, grey nodes indicate known values, white nodes indicate hidden variables.

The probability under the model of a french sentence **f** given an English sentence **e** thus becomes:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \prod_{j=1}^{J} p^{*}(f_j|e_{a_j}) p(a_j|a_{j-1}, I) \quad (2)$$

where

$$p^{*}(f_j|e_{a_j}) = \sum_{k=1}^{K} p(f_j|s_k) p(s_k|e_{a_j}) \quad (3)$$

Here, K is the number of senses that english word associated with this translation pair. The senses will be constrained either by the English word e_{a_j} or by the French word f_j depending on which language the sense inventory is taken from. The first case, with senses constrained by the English, will be denoted with SHMM1, and the second with SHMM2. In this work, only SHMM1 is used.

If the amount of senses defined for each word is exactly 1 and this sense is different for each word, the model reduces to the HMM model (see Figure 2). However, if the sense inventory is defined such that for two different words with a sense that is similar, the same sense can be used, the model is able to use translation probabilities drawn from observations from both these words together. For example, in SHMM1, the words ‘small’ and ‘little’ may have the same sense listed in the sense inventory, which allows the model to learn a translation distribution to the French words that both these words often align to.

For training this model, as with the IBM models, Expectation-Maximization and initialisation are key. The more complex IBM models are initialised from simpler versions, so the complex models can start out with reasonable estimates, which allow it to find good alignments. Here, too, the same steps are used. The HMM model is initialised from Model 1, as described in citevogel:1996. From this, the SHMM models can be initialised.

For the SHMM1, given a translation probability for a french word given an english word under the HMM, $p(f|e)$, and a list of valid senses for that english word e, an equal portion of that translation probability is given to the new translation probability depending on the sense. This is done for all translation probabilities, and the translation table is then normalised. Probability of a sense given an english word is initialised to a uniform distribution over the valid senses.

For the SHMM2, the probability of french words given a sense is set to uniform over the words for which the sense is valid, and the probability of the sense given the english word is calculated analogous to the probability of the french word given the sense in the first case.

After initialisation, the expectation-maximisation algorithm can be used for training, as with the HMM model, using the forward-backward algorithm to find the posterior probabilities of the alignments. As the senses can be summed out during this phase, the algorithm can be used as-is, and afterwards the proportion of the partial count that should be assigned to each sense can be found. By summing out over the relevant senses and words, the two parts $c(f_j|q_k)$ and $c(q_k|e_i)$ can then be found.

3.1 Generating Senses for Words

In order to be able to use this model, an inventory of senses is needed for every word in the corpus, for one of the languages. The most obvious source for this is the English Wordnet (Miller, 1995), as it has a large inventory of senses. Note that, in this document, the words senses and synsets are used interchangeably.

The process of obtaining this inventory is explained from the viewpoint of using English WordNet, but the same basic conditions apply for any other lexicon, or language. The inventory of senses is obtained through the WordNet corpus in NLTK [1], which automatically stems the words that synsets are sought for.

In this model, two senses (synsets) are functionally equivalent, if the list of words that have them in their senselist is the same for both senses. That is to say, if the partial counts that will be added to either of the senses will be the same, there is no way of distinguishing between the two senses under this model. For example, in WordNet 3.0, among the synsets listed for the word ‘small’, there are 3 that have as constituent words only ‘small’ and ‘little’. These 3 synsets would be functionally equivalent for our purposes. When this occurs, the senses that are equivalent are collated under one name, so that it’s possible to find out which senses a particular sense is made up of.

At this point, there will be some words with only a sense that is unique to that word (such as those words that were not in the lexicon, which get a newly made sense), some words with only shared senses and some with a mix. We might want to enforce one of a few distinct options:

- All words have exactly 1 unique sense, and perhaps a few shared ones (‘synthesis’ condition)
- Some words have a unique sense, some don’t (‘merge’ condition)
- No words have unique senses if they have at least 1 shared sense (‘none’ condition)

These conditions are generated by first finding the filtered list of senses for each word. At this point, some words have only unique senses, either because they didn’t occur in WordNet, or because WordNet only listed unique senses for that word (the ‘merge’ condition. The ‘synth’ condition is made, by finding all words that have only shared senses, and adding a new sense, that is unique to that word. The ‘none’

[1] http://www.nltk.org/

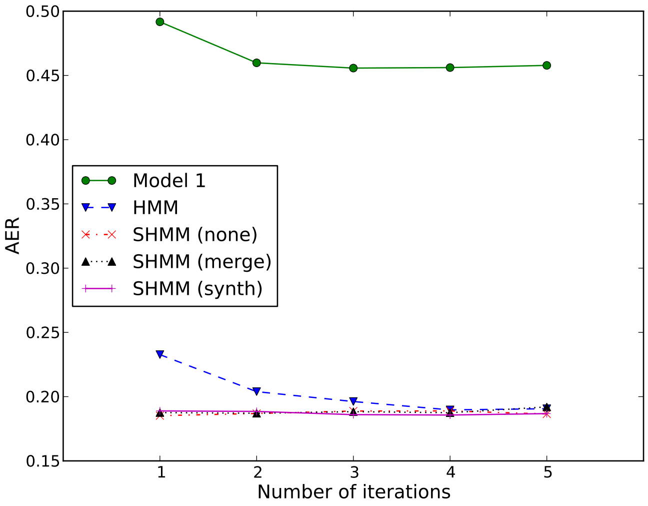

Figure 3: AER scores for Model 1, HMM, and 3 SHMM variations trained for 5 iterations each, lower is better.

condition then is found by doing the opposite: removing all unique senses from words that also have shared senses.

Under each of these 3 conditions, the model might work slightly differently. Under the 'synthesis' condition, it may generate the translation probabilities either directly, as in the HMM (which is what happens for any word with only 1 sense, which is unique for that word), or from the shared probabilities, through the senses. In the other models, the model is increasingly forced to use the shared translation probabilities.

4 Evaluation

We will evaluate the early results of this model against the HMM and Model 1 results, and will do a qualitative analysis of the distribution over senses and French words that the model obtains, in order to find out if reasonable predictions for senses are made.

The sense HMM model will be evaluated using the three sense inventories suggested in subsection 3.1. The dataset used was a 1 million sentence aligned English-French corpus, taken from the Europarl corpus (Koehn, 2005). The data was tokenised, length limited to a maximum length of 50, and lowercased. The results are evaluated on the test set from the ACL 2005 shared task, using Alignment Error Rate. The models are all trained for 5 iterations, and a pruning threshold is employed that removes probabilities from the translation tables if it is below $1.0 \cdot 10^{-6}$.

The results of training models based on senses generated in the 3 ways listed above is shown in Figure 3. The three SHMM models are compared against Model 1, and the standard HMM model, each of which is trained for 5 iterations. The HMM model is initialised from Model 1, and the SHMM models initialised from the HMM model. As the figure shows, the AER score for the last two iterations of the HMM model is very similar to the scores that the three variations of the SHMM model attain. The scores for the three HMM models range from 0.185 to 0.192

A possible reason for this performance is that the models didn't have enough sharing going on between the senses. The corpus contains 70700 unique words. Looking at the amount of senses that are found in the 'none' condition, meaning that all of the WordNet senses share output probabilities, there are 17194 words that have at least one of these senses listed, and there are 27120 distinct senses available in that setting. For the other 53500 senses, no sharing is going on whatsoever.

In the 'merge' and 'synth' conditions, there are more senses taken from WordNet (for a total from WordNet of 33133), but these don't add any shar-

Sense	Definition	$P(s\|e)$	Most likely French words in order
severe.s.06	very bad in degree or extent	0.4861	graves, sévères, des, sévère, grave, de, gravement, une, sérieuses, les
severe.s.04	unsparing and uncompromising in discipline or judgment	0.2358	graves, sévères, des, sévère, grave, de, gravement, une, sérieuses, les
dangerous.s.02	causing fear or anxiety by threatening great harm	0.1177	grave, des, graves, les, sérieux, très, sérieuses, une, importantes, sérieuse
austere.s.01	severely simple	0.1148	graves, des, grave, sévère, sévères, très, fortement, forte, rigoureuses, situation
hard.s.04	very strong or vigorous	0.035	dur, plus, importants, des, sévères, durement, son, une, difficile, très
severe.s.01	intensely or extremely bad or unpleasant in degree or quality	0.01055	terrible, terribles, des, grave, les, mauvais, dramatique, cette, aussi, terriblement

Table 1: Senses for the word ‘severe’ in the ‘none’ version of the SHMM model, their WordNet definition, the probability of the sense for the word severe, and the most likely French words for the senses given in order of likelihood.

ing. It might be then, that the model has insufficient opportunity to share output distributions, causing it to behave much as the HMM alignment model. Another possibility is, that the senses insufficiently well-defined, and share probabilities between words that are too dissimilar, negating any positive effect this may have and possibly pushing the model towards less sharing. We will suggest possibilities for dealing with this in section 5.

Regardless of the performance of the model in word alignment, if the model learns probabilities for senses that are reasonable, it can be used as a word sense disambiguation system for parallel corpora, with the candidate senses being made up from the senses out of WordNet. Those words not listed in WordNet, are treated as being monosemous words in this context. The ‘merge’ and ‘none’ conditions are most useful for this: if a WSD system chooses a sense that is not linked to a WordNet sense, it is not clearly defined which sense is meant here.

In order to find out if the model makes sensible distinctions between different senses, we have picked a random polysemous word, and looked at the senses associated with it in the ‘none’ condition. The word that was chosen is ‘severe’. It has 6 possible senses, listed by main word and definition in Table 1, along with the probability of the senses, $p(s|e)$, and the 10 most likely French words for the senses.

Sense	Associated English words
severe.s.06	(only has basic 3 senses)
severe.s.04	spartan
dangerous.s.02	dangerous, grave, graver, gravest, grievous, life-threatening, serious
austere.s.01	austere, stark, starker, starkest, stern
hard.s.04	hard, harder, hardest
severe.s.01	terrible, wicked

Table 2: Senses for the word ‘severe’ in the ‘none’ version of the SHMM model and the English words apart from ‘severe’, ‘severer’ and ‘severest’ that have the sense in their senselist

As the table shows, the two most likely senses are quite similar. In fact, because words are stemmed before looking up suitable senses, all senses have at least the following 3 words associated with them: ‘severe’, ‘severer’ and ‘severest’. The words that

Sense	Definition	P(s—e)	Most likely French words in order
rigorous.s.01	rigidly accurate; allowing no deviation from a standard	0.8962	rigoureuse, rigoureux, une, rigueur, rigoureuses, des, un, stricte, strict, strictes
rigorous.s.02	demanding strict attention to rules and procedures	0.1038	des, strictes, rigoureux, stricte, sévères, rigoureuses, stricts, rigoureuse, une, sévère

Table 3: Senses for the word 'rigorous' in the 'none' version of the SHMM model, their WordNet definition, the probability of the senses of the word 'rigorous', and the most likely French words for the senses given in order of likelihood.

cause the differences between the senses are listed in table 2. It can be seen that the only difference between severe.s.04 and severe.s.06 is the addition of the word 'spartan' for the first. As 'spartan' only occurs 67 times in the corpus, versus 484 for severe, it is possible that they are so similar, because the counts for 'spartan' get overshadowed.

For the other senses however, the most likely translations vary quite a bit. The sense 'hard.s.04', meaning very strong or vigorous, also includes translations to 'plus' and 'dur', which seems more likely given the sense. Given these translation probabilities though, it should at least be possible to distinguish between different senses of the word severe, given that it's aligned to a different french word.

One more example is listed in table 3, showing the probabilities for two different senses, and their most likely translations. The most likely sense for rigorous under the model is in the sense of 'allowing no deviation from a standard'. This is the only of the two senses that can translate to 'rigueur' in french, literally rigor. The other sense, meaning 'demanding strict attention to rules and procedures', is more likely to translate to 'strictes', 'stricte' and 'sévères', which reflects the WordNet definition.

The difference in contributing English words between these two senses can be found in Table 4. Interestingly, the three forms of the word strict are associated with the sense rigorous.s.01, even though the naive translations of these words into French are more likely for rigorous.s.02. Even so, the results match the WordNet definitions better.

Sense	Associated English words
rigorous.s.01	rigorous strict stricter strictest
rigorous.s.02	rigorous stringent tight tighter tightest

These results show that useful translations are found, and the corresponding senses can be learned as well. For sense discrimination in parallel corpuses then, this model shows potential, and for alignment good alignments can be found, even with better abstraction in the model.

5 Conclusion

The results have shown that this may be a useful way to incorporate senses in a word alignment system. While the alignment results in themselves weren't significantly better, alignment probabilities to senses have been shown to be generated, which make it possible to distinguish between different senses. This could open the door to automatically sense annotating parallel corpora, using a predefined set of senses.

At this early point, several options lay open to improve upon the results so far. To improve the alignment results, more encompassing senses may be generated, for example by integrating similar synsets. At the same time, the list of synsets for each word may be improved upon, by filtering out very unlikely senses for a word.

It should also be possible to employ an already existing WSD system to annotate the parallel corpus, and use the counts of the annotated senses to better initialise the senses, rather than starting out assuming all are equaly likely for a given word. This may be used as well to initialise the translation probabilities for senses.

References

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 387–394, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *In The 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007*, pages 61–72.

Yee Seng Chan and Hwee Tou Ng. 2007. Word sense disambiguation improves statistical machine translation. In *In 45th Annual Meeting of the Association for Computational Linguistics (ACL-07*, pages 33–40.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA 2004 (Conference of the Association for Machine Translation in the Americas)*, volume 3265, pages 115–124. Springer.

P. Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38:39–41.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 455–462, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Tufiş, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.