

Extensions to HMM-based Statistical Word Alignment Models

Kristina Toutanova, H. Tolga Ilhan and Christopher D. Manning

Department of Computer Science

Stanford University

Stanford, CA 94305-9040 USA

kristina@cs.stanford.edu

ilhan@stanford.edu

manning@cs.stanford.edu

Abstract

This paper describes improved HMM-based word level alignment models for statistical machine translation. We present a method for using part of speech tag information to improve alignment accuracy, and an approach to modeling fertility and correspondence to the empty word in an HMM alignment model. We present accuracy results from evaluating Viterbi alignments against human-judged alignments on the Canadian Hansards corpus, as compared to a bigram HMM, and IBM model 4. The results show up to 16% alignment error reduction.

1 Introduction

The main task in statistical machine translation is to model the string translation probability $P(e_1^l | f_1^m)$ where the string f_1^m in one language is translated into another language as string e_1^l . We refer to e_1^l as the source language string and f_1^m as the target language string in accordance with the noisy channel terminology used in the IBM models of (Brown et al., 1993). Word-level translation models assume a pairwise mapping between the words of the source and target strings. This mapping is generated by alignment models. In this paper we present extensions to the HMM alignment model of (Vogel et al., 1996; Och and Ney, 2000b). Some of our extensions are applicable to other alignment models as well and are of general utility.¹

For most language pairs huge amounts of parallel corpora are not readily available whereas monolingual resources such as taggers are more often available. Little research has gone into exploring the po-

tential of part of speech information to better model translation probabilities and permutation probabilities. Melamed (2000) uses a very broad classification of words (content, function and several punctuation classes) to estimate class-specific parameters for translation models. Fung and Wu (1995) adapt English tags for Chinese language modeling using Coerced Markov Models. They use English POS classes as states of the Markov Model to generate Chinese language words. In this paper we use POS tag information to incorporate prior knowledge of word translation and to model local word order variation. We show that using this information can help in the translation modeling task.

Many alignment models assume a one to many mapping from source language words to target language words, such as the IBM models 1-5 of Brown et al. (1993) and the HMM alignment model of (Vogel et al., 1996). In addition, the IBM Models 3, 4 and 5 include a fertility model $p(\sigma|e)$ where σ is the number of words aligned to a source word e . In HMM-based alignment word fertilities are not modeled. The alignment positions of target words are the states in an HMM. The alignment probabilities for word f_j depend only on the alignment of the previous word f_{j-1} if using a first order HMM. Therefore, source words are not awarded/penalized for being aligned to more than one target word. We present an extension to HMM alignment that approximately models word fertility.

Another assumption of existing alignment models is that there is a special Null word in the source sentence from which all target words that do not have other correspondences in the source language are generated. Use of such a Null word has proven problematic in many models. We also assume the

¹This paper was supported in part by the National Science Foundation under Grants IIS-0085896 and IIS-9982226. The authors would also like to thank the various reviewers for their helpful comments on earlier versions.

existence of a special Null word in the source language that generates words in the target language. However, we define a different model that better constrains and conditions generation from Null. We assume that the generation probability of words by Null depends on other words in the target sentence.

Next we present the general equations for decomposition of the translation probability using part of speech tags and later we will go into more detail of our extensions.

2 Part of Speech Tags in a Translation Model

Augmenting the model $P(e_1^l | f_1^m)$ with part of speech tag information leads to the following equations. We use e_1^l , f_1^m or vector notation \mathbf{e} , \mathbf{f} to denote English and French strings. (l and m represent the lengths of the French and English strings respectively.) Let us define \mathbf{eT} and \mathbf{fT} as possible POS tag sequences of the sentences \mathbf{e} and \mathbf{f} . We can rewrite the string translation probability $P(e_1^l | f_1^m)$ as follows (using Bayes rule to give the last line):

$$\begin{aligned} \mathbf{P}(\mathbf{e}|\mathbf{f}) &= \sum_{\mathbf{fT}} \mathbf{P}(\mathbf{e}, \mathbf{fT}|\mathbf{f}) \\ &= \sum_{\mathbf{fT}} \mathbf{P}(\mathbf{fT}|\mathbf{f})\mathbf{P}(\mathbf{e}|\mathbf{f}, \mathbf{fT}) \\ &= \sum_{\mathbf{fT}} \mathbf{P}(\mathbf{fT}|\mathbf{f}) \sum_{\mathbf{eT}} \mathbf{P}(\mathbf{e}, \mathbf{eT}|\mathbf{f}, \mathbf{fT}) \\ &= \sum_{\mathbf{fT}} \mathbf{P}(\mathbf{fT}|\mathbf{f}) \sum_{\mathbf{eT}} \mathbf{P}(\mathbf{e})\mathbf{P}(\mathbf{eT}|\mathbf{e}) \frac{\mathbf{P}(\mathbf{f}, \mathbf{fT}|\mathbf{e}, \mathbf{eT})}{\mathbf{P}(\mathbf{f}, \mathbf{fT})} \end{aligned}$$

If we also assume that the taggers in both languages generate a single tag sequence for each sentence then the equation for machine translation by the noisy channel model simplifies to

$$\arg \max_{\mathbf{e}} \mathbf{P}(\mathbf{e}|\mathbf{f}) = \arg \max_{\mathbf{e}} \mathbf{P}(\mathbf{e})\mathbf{P}(\mathbf{f}, \mathbf{fT}|\mathbf{e}, \mathbf{eT})$$

This is the decomposition of the string translation probability into a language model and translation model. In this paper we only address the translation model and assume that there exists a one-to-one alignment from target to source words. Therefore,

$$\mathbf{P}(\mathbf{f}, \mathbf{fT}|\mathbf{e}, \mathbf{eT}) = \sum_{\mathbf{a}} \mathbf{P}(\mathbf{f}, \mathbf{fT}, \mathbf{a}|\mathbf{e}, \mathbf{eT})$$

One possible way to rewrite $\mathbf{P}(f_1^m, fT_1^m, a_1^m | e_1^l, eT_1^l)$ without loss of generality is:

$$\mathbf{P}(f_1^m, fT_1^m, a_1^m | \mathbf{e}, \mathbf{eT}) = \mathbf{P}(m | \mathbf{e}, \mathbf{eT}) \times \prod_{j=1}^J \left\{ \begin{array}{l} \mathbf{P}(a_j | a_1^{j-1}, f_1^{j-1}, fT_1^{j-1}, m, \mathbf{e}, \mathbf{eT}) \\ \mathbf{P}(fT_j | a_1^j, f_1^{j-1}, fT_1^{j-1}, m, \mathbf{e}, \mathbf{eT}) \\ \mathbf{P}(f_j | a_1^j, f_1^{j-1}, fT_1^j, m, \mathbf{e}, \mathbf{eT}) \end{array} \right\} \quad (1)$$

Here each a_j gives the index of the word e_{a_j} to which f_j is aligned. The models we present in this paper will differ in the decompositions of alignment probabilities, tag translation and word translation probabilities in Eqn. 1. Section 3 describes the baseline model in more detail. Section 4 illustrates examples where the baseline model performs poorly. Section 5 presents our extensions and Section 6 presents experimental results.

3 Baseline Model

Translation of French and English sentences shows a strong localization effect. Words close to each other in the source language remain close in the translation. Furthermore, most of the time the alignment shows monotonicity. This means that pairwise alignments stay close to the diagonal line of the (j, i) plane. It has been shown (Vogel et al., 1996; Och et al., 1999; Och and Ney, 2000a) that HMM based alignment models are effective at capturing such localization.

We use as a baseline the model presented by (Och and Ney, 2000a). A basic bigram HMM-based model gives us

$$\mathbf{P}(\mathbf{f}|\mathbf{e}) = \sum_{a_1^m} \prod_{j=1}^m \left[\mathbf{P}(a_j | a_{j-1}, l) \mathbf{P}(f_j | e_{a_j}) \right] \quad (2)$$

In this HMM model,² alignment probabilities are independent of word position and depend only on jump width $(a_j - a_{j-1})$.³ The Och and Ney (2000a) model includes refinements including special treatment of a jump to Null and smoothing with a uniform prior which we also included in our initial model. As in their model we set the probability for jump from any state to Null to a fixed value ($p = .4$) which we estimated from held-out data.

²Each HMM state is $[a_j, e_{a_j}]$ emitting f_j as output.

³In order for the model not to be deficient, we normalize the jump probabilities at each EM step so that jumping outside of the borders of the sentence is not possible.

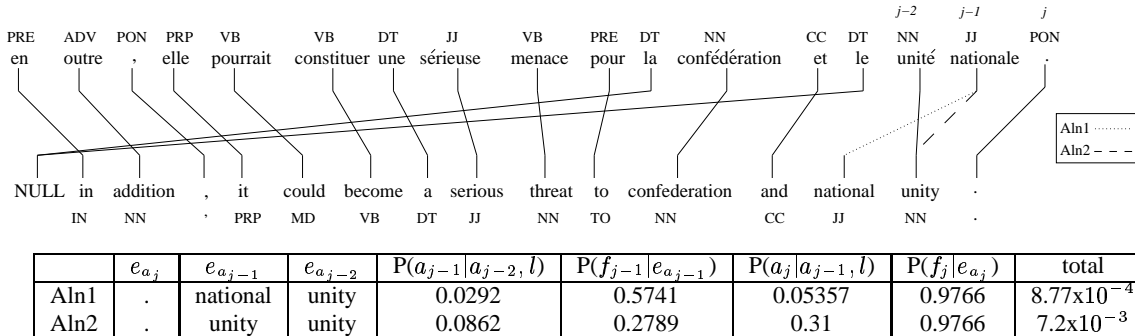


Figure 1: The baseline model makes a simple alignment error.

4 Alignment Irregularities

Although the baseline Hidden Markov alignment model successfully generates smooth alignments, there are a fair number of alignment examples where pairwise match shows local irregularities. One instance of this is the transition of the NP \rightarrow JJ NN rule to NP \rightarrow NN JJ from English to French. We can list two main reasons why word translation probabilities may not catch such irregularities to monotonicity. First, it may be the case that both the English adjective and noun are words that are unknown. In this case the translation probabilities will be close to each other after smoothing. Second, the adjective-noun pair may consist of words that are frequently seen together in English. *National reserve* and *Canadian parliament*, are examples of such pairs. As a result there will be an indirect association between the English noun and the translation of the English adjective. In both cases, word translation probabilities will not be differentiating enough and alignment probabilities become the dominating factor to determine where f_j aligns.

Figure 1 illustrates how our baseline HMM model makes an alignment mistake of this sort. The table in the figure displays alignment and translation probabilities of two competing alignments (namely *Aln1* and *Aln2*) for the last three words. In both alignments, the shown f_j and e_{a_j} are periods at the end of the French and English sentences. The first alignment maps *nationale* to *national* and *unité* to *unity*. (i.e. $e_{a_{j-1}} = \textit{national}$ and $e_{a_{j-2}} = \textit{unity}$). The second alignment maps both *nationale* and *unité* to *unity* (i.e. $e_{a_{j-1}} = \textit{unity}$ and $e_{a_{j-2}} = \textit{unity}$). Starting from the *unity-unity* alignment, the jump width

sequences $\{(a_{j-1} - a_{j-2}), (a_j - a_{j-1})\}$ for *Aln1* and *Aln2* are $\{-1, 2\}$ and $\{0, 1\}$ respectively. The table shows that the gain from use of monotonic alignment probabilities dominates over the lowered word translation probability. Although *national* and *nationale* are strongly correlated according to the translation probabilities, jump widths of -1 and 2 are less probable than jump widths of 0 and 1 .

5 Extensions

In this section we describe our improvements on the HMM model. We present evaluation results in Section 6 after describing the technical details of our models here.

5.1 POS Tags for Translation Probabilities

Our model with part of speech tags for translation probabilities uses the following simplification of the translation probability shown in Eqn. 1.⁴

$$\mathbf{P}(\mathbf{f}, \mathbf{fT} | \mathbf{e}, \mathbf{eT}) = \sum_{a_1^m} \prod_{j=1}^m \left\{ \begin{array}{l} \mathbf{P}(a_j | a_{j-1}, l) \\ \mathbf{P}(f_j | e_{a_j}) \\ \mathbf{P}(f_j | e_{a_j}) \end{array} \right\} \quad (3)$$

In this model we introduce tag translation probabilities as an extra factor to Eqn. 2. Intuitively the role of this factor is to boost the translation probabilities for words of parts of speech that can often be translations of each other. Thus this probability distribution provides prior knowledge of the possible translations of a word based only on its part of speech. However, $\mathbf{P}(f_j | e_{a_j})$ should not be too sharp or

⁴Since we are only concerned with alignment here and not generation of candidate translations the factor $\mathbf{P}(m | \mathbf{e}, \mathbf{eT})$ can be ignored and we omit it from the equations for the rest of the paper.

it will dominate the alignment probabilities and the probabilities $\mathbf{P}(f|e)$. We use the following linear interpolation to smooth tag translation probabilities:

$$\mathbf{P}(fT_j|eT_{a_j}) = \lambda \tilde{\mathbf{P}}(fT_j|eT_{a_j}) + (1 - \lambda) \frac{1}{T} \quad (4)$$

T is the size of the French tag set and λ is set to be 0.1 in our experiments. The tag translation model is so heavily smoothed with a uniform distribution because in EM the tag translation probabilities quickly become very sharp and can easily overrule the alignment and word translation probabilities. The Results section shows that the addition of this factor reduces the alignment error rate, with the improvement being especially large when the training data size is small.

5.2 Tag Sequences for Jump Probabilities

This section describes an extension to the bigram HMM model that uses source and target language tag sequences as conditioning information when predicting the alignment of target language words.

In the decomposition of the joint probability $\mathbf{P}(\mathbf{f}, \mathbf{fT}, \mathbf{a}|\mathbf{e}, \mathbf{eT})$ shown in Eqn. 1 the factor for alignment probabilities is $\mathbf{P}(a_j|a_1^{j-1}, f_1^{j-1}, fT_1^{j-1}, m, \mathbf{e}, \mathbf{eT})$.

A bigram HMM model assumes independence of a_j from anything but the previous alignment position a_{j-1} and the length of the English sentence. Brown et al. (1993) and Och et al. (1999) variably condition this probability on the English word in position a_{j-1} and/or the French word in position j . As conditioning directly on words would yield a large number of parameters and would be impractical, they cluster the words automatically into bilingual word classes.

The question arises then whether we would have larger gains by conditioning on the part of speech tags of those words or even more words around the alignment position. For example, if we use the following conditioning information:

$$\mathbf{P}(a_j|a_1^{j-1}, f_1^{j-1}, fT_1^{j-1}, m, \mathbf{e}, \mathbf{eT}) = \mathbf{P}(a_j|a_{j-1}, eT_{a_{j-1}-1}, eT_{a_{j-1}}, eT_{a_{j-1}+1})$$

we could model probabilities of transpositions and insertion of function words in the target language that have no corresponding words in the source language (e_{a_j} is Null) similarly to the channel operations of the (Yamada and Knight, 2001) syntax-

based statistical translation model. Since the syntactic knowledge provided by POS tags is quite limited, this is a crude model of transpositions and Null insertions at the preterminal level. However we could still expect that it would help in modeling local word order variations. For example, in the sentence *J'aime la chute* 'I love the fall' the probability of aligning $f_j = \text{la}$ ($fT_j = DT$) to **the** will be boosted by knowing $eT_{a_{j-1}} = VBP$ and $eT_{a_{j-1}+1} = DT$. Similarly, in the sentence *J'aime des chiens* 'I love dogs' the probability of aligning $f_j = \text{la}$ to Null will be increased by knowing $eT_{a_{j-1}} = VBP$ and $eT_{a_{j-1}+1} = NNS$. *VBP* followed by *NNS* crudely conducts the information that the verb is followed by a noun phrase which does not include a determiner.

We conducted a series of experiments where the alignment probabilities are conditioned on different subsets of the part of speech tags $eT_{a_{j-1}-1}, eT_{a_{j-1}}, eT_{a_{j-1}+1}, fT_{j-1}, fT_j, fT_{j+1}$.

In order to be able to condition on fT_j, fT_{j+1} when generating an alignment position for f_j , we have to change the generative model for the sentence \mathbf{f} and its tag sequence \mathbf{fT} to generate the part of speech tags for the French words before choosing alignment positions for them. The French POS tags could be generated for example from a prior distribution $\mathbf{P}(fT_j)$ or from the previous French tags as in an HMM for part-of-speech tagging. The generative model becomes: $\mathbf{P}(\mathbf{f}, \mathbf{fT}|\mathbf{e}) = \mathbf{P}(\mathbf{fT}) \sum_{a_1^m} \prod_{j=1}^m [\mathbf{P}(a_j|a_{j-1}, fT_j, fT_{j+1}, l) \mathbf{P}(f_j|e_{a_j})]$

This model makes the assumption that target words are independent of their tags given the corresponding source word and models only the dependence of alignment positions on part of speech tags.

5.3 Modeling Fertility

A major advantage of the IBM models 3–5 over the HMM alignment model is the presence of a model of source word fertility. Thus knowledge that some words translate as phrases in the target language is incorporated in the model.

The HMM model has no memory, apart from the previous alignment, about how many words it has aligned to a source word. Yet even this memory is not used to decide whether to generate more words from a given English word. The decision to gener-

ate again (to make a jump of size 0) is independent of the word and is estimated over all words in the corpus.

We extended the HMM model to decide whether to generate more words from the previous English word $e_{a_{j-1}}$ or to move on to a different word depending on the identity of the English word $e_{a_{j-1}}$. We introduced a factor $\mathbf{P}(\text{stay}|e_{a_{j-1}})$ where the boolean random variable `stay` depends on the English word f_{j-1} aligned to. Since in most cases words with fertility greater than one generate words that are consecutive in the target language, this extension approximates fertility modeling. More specifically, the baseline model (i.e., Eqn. 2) is changed as follows:

$$\mathbf{P}(\mathbf{f}|\mathbf{e}) = \sum_{a_1}^m \prod_{j=1}^m \left[\tilde{\mathbf{P}}(a_j|a_{j-1}, e_{a_{j-1}}, l) \mathbf{P}(f_j|e_{a_j}) \right]$$

where

$$\tilde{\mathbf{P}}(a_j|a_{j-1}, e_{a_{j-1}}, l) = \delta(a_j, a_{j-1}) \mathbf{P}(\text{stay}|e_{a_{j-1}}) + \left\{ \begin{array}{l} (1 - \delta(a_j, a_{j-1})) \\ (1 - \mathbf{P}(\text{stay}|e_{a_{j-1}})) \\ \mathbf{P}(a_j|a_{j-1}, l) \end{array} \right\} \quad (5)$$

$\delta(a_j|a_{j-1})$ in Eqn. 5 is the Kronecker delta function. Basically, the new alignment probabilities $\tilde{\mathbf{P}}(a_j|a_{j-1}, l)$ state that a jump width of zero depends on the English word. If we define the fertility of a word as the number of consecutive words from the target language it generates, then the probability distribution for the fertility of an English word e according to this model is geometric with a probability of success $1 - \mathbf{P}(\text{stay}|e)$. The expectation is $\frac{1}{1 - \mathbf{P}(\text{stay}|e)}$.⁵ Even though the fit of this distribution to the real fertility distribution may not be very good, this approximation improves alignment accuracy in practice.

Sparsity is a problem in estimating `stay` probabilities $\mathbf{P}(\text{stay}|e_{a_{j-1}})$. We use the probability of a jump of size zero from the baseline model as our prior to do smoothing as follows:

$$\mathbf{P}(\text{stay}|e_{a_{j-1}}) = \lambda \mathbf{P}_{ZJ} + (1 - \lambda) \mathbf{P}(\text{stay}|e_{a_{j-1}}) \quad (6)$$

⁵ $\mathbf{E}[X] = \frac{1}{p} = p + 2p(1-p) + 3p(1-p)^2 + \dots$ where X is the number of Bernoulli trials until the first success.

\mathbf{P}_{ZJ} in this equation is the alignment probability from the baseline model with zero jump distance.

$$\mathbf{P}_{ZJ} = \mathbf{P}(a_j = i | a_{j-1} = i, l).$$

5.4 Translation Model for Null

As originally proposed by Brown et al. (1993), words in the target sentence for which there are no corresponding English words are assumed to be generated by the special English word `Null`. `Null` appears in every English sentence and often serves to generate syntactic elements in the target language that are missing in the source. A probability distribution $\mathbf{P}(f|\text{Null})$ for generation probabilities of the `Null` is re-estimated from a training corpus.

Modeling a `Null` word has proven problematic. It has required many special fixes to keep models from aligning everything to `Null` or to keep them from aligning nothing to `Null` (Och and Ney, 2000b). This might stem from the problem that the `Null` is responsible for generating syntactic elements of the target language as well as generating words that make the target language sentence more idiomatic and stylistic. The intuition for our model of translation probabilities for target words that do not have corresponding source words is that these words are generated from the special English `Null` and also from the next word in the target language by a mixture model. The pair *la confédération* in Figure 1 is an example of such case where *confédération* contributes extra information in generation of *la*. The formula for the probability of a target word given that it does not have a corresponding aligning word in the source is:

$$\mathbf{P}(f_j|a_j = 0) = \lambda \mathbf{P}(f_j|f_{j+1}, e_{a_j} = \text{Null}) + (1 - \lambda) \mathbf{P}(f_j|e_{a_j} = \text{Null}) \quad (7)$$

We re-estimate the probabilities $\mathbf{P}(f_j|f_{j+1}, e_{a_j} = \text{Null})$ from the training corpus using EM. The dependence of a French word on the next French word requires a change in the generative model to first propose alignments for all words in the French sentence and to then generate the French words given their alignments, starting from the end of the sentence and going towards the beginning. For the new model there is an efficient dynamic programming algorithm for computations in EM similar to the forward-backward algorithm. The probability

$\mathbf{P}(f_1 \dots f_j, a_j = i, f_{j+1}, \dots f_m | \mathbf{e})$ again decomposes into forward and backward probabilities. The forward probability is $\alpha(j, i) = \mathbf{P}(a_j = i) \times \mathbf{P}(f_1 \dots f_j | a_j = i, f_{j+1}, \mathbf{e})$ and the backward probability is $\beta(j, i) = \mathbf{P}(f_{j+1} \dots f_m | a_j = i, \mathbf{e})$. These can be computed recursively and used for efficient computation of posteriors in EM.

6 Results

We present results on word level alignment accuracy using the Hansards corpus. Our test data consists of 500 manually aligned sentences which are the same data set used by (Och and Ney, 2000b).⁶ In the annotated sentences every alignment between two words is labeled as either a sure (S) or possible (P) alignment. ($S \subseteq P$). We used the following quantity (called alignment error rate or AER) to evaluate the alignment quality of our models, which is also the evaluation metric used by (Och and Ney, 2000b):

$$\text{recall} = \frac{|A \cap S|}{|S|} \quad \text{precision} = \frac{|A \cap P|}{|A|}$$

$$\text{AER} = \frac{|A \cap P| + |A \cap S|}{|A| + |S|}$$

We divided this annotated data into a validation set of 100 sentences and a final test set of 400 sentences. The validation set was used to select tuning parameters such as λ in Eqn. 4, 6 and 7. We report AER results on the final test set of 400 sentences which contain a total of 6,365 English and 7,018 French words. We experimented with training corpora of different sizes ranging from 5K to 50K sentences. We concentrated on small to medium data sets to assess the ability of our models to deal with sparse data.

Table 1 shows the percentage of words in the corpus that were seen less than the specified number of times. For example, in our 10K training corpus 47% of all word types were seen only once. As seen from the table the sparsity is great even for large corpora.

The models we implemented and compare in this section are the following:

- **Baseline** is the baseline HMM model described in section 2
- **Tags** is an HMM model that includes tags for translation probabilities (section 5.1)

⁶We want to thank Franz Och for sharing the annotated data with us.

- **SG** is an HMM model that includes *stay* probabilities (section 5.3)
- **Null** is an HMM model that includes the new generation model for words by Null (section 5.4)
- **Tags+Null**, **Tags+SG**, and **Tags+Null+SG** are combinations of the above models

Table 2 shows AER results for our improved models on training corpora of increasing size. The model **Null** outperforms the baseline at every data set size, with the error reduction being larger for bigger training sets (up to 9.2% error reduction). The **SG** model reduces the baseline error rate by up to 10%. The model **Tags** reduces the error rate for the smallest dataset by 7.6%. The combination of **Tags** and the **SG** or **Null** models outperforms the individual models in the combination since they address different problems and make orthogonal mistakes. The combination of **SG** and **Tags** reduces the baseline error rate by up to 16% and the combination of **Null** and **Tags** reduces the error rate by up to 12.3%. All of these error reductions are statistically significant at the $\sigma = .05$ confidence level according to the paired t-test. The combination **Tags+Null+SG** further reduces the error rate. For small datasets, there seems to be a stronger overlap between the strengths of the **Null** and **SG** models because some fertility related phenomena can be accounted for by both models. When an English word is wrongly aligning to several consecutive French words because of indirect association, while the correct alignment of some of them is to the empty word, both the **Null** and **SG** models can combat the problem— one by better modeling correspondence to **Null**, and the other by discouraging large fertilities.

Figure 2 displays learning curves for three models: **Och**, **Tags**, and **Tags+Null**. **Och** is the HMM alignment model of (Och and Ney, 2000b). To obtain results from the **Och** model we ran GIZA++.⁷ Both the **Tags** and **Och** models use word classes. However the word classes used in the latter are learned automatically from parallel bilingual corpora while the classes used in the former are human defined part of speech tags. Figure 2 shows that the **Tags** model outperforms the **Och** model when the training data size is small. As the train-

⁷GIZA++ can be downloaded from <http://www-i6.informatik.rwth-aachen.de/~och/software/GIZA++.html>

Table 1: Percentage of words in the corpus by frequency

Corpus	= 1		< 3		< 5		< 10	
	English	French	English	French	English	French	English	French
10K	47%	50%	61%	66%	74%	77%	84%	87%
25K	43%	44%	57%	59%	69%	72%	80%	83%
50K	42%	44%	55%	57%	67%	69%	78%	81%

Table 2: Alignment Error Rate by Model and Corpus Size

Corpus	Baseline	Null	SG	Tags	Tags+SG	Tags+Null	Tags+Null+SG
5K	17.53	16.86	16.72	16.20	15.31	15.36	15.14
15K	15.03	14.29	13.52	13.90	12.63	13.22	12.52
25K	13.85	13.05	12.79	13.10	11.91	12.30	11.79
35K	13.19	11.98	12.03	12.60	11.45	11.56	11.07
50K	12.63	11.76	11.78	12.10	11.19	11.11	10.69

ing size increases the Och model catches up with the Tags model and even surpasses it slightly. This suggests that when large amounts of parallel text are not available monolingual part of speech classes can improve alignment quality more than automatically induced classes. When more data is available automatically induced bilingual word classes seem to provide more improvement but it still remains to be explored whether the combination of part-of-speech knowledge with induction of bilingual classes will perform even better. The third curve in the figure for Tags+Null illustrates the relative improvement of the Null model over the Tags model as the training set size increases. We see that the performance gap between the two models becomes wider for larger training data sizes. This reflects the improved estimation of the generation probabilities for Null which require target word specific parameters. We used

both paired t-test and Wilcoxon signed rank tests to show the improvements are statistically significant. The signed rank test uses the normalized test statistic $\frac{W_+ - E(W_+)}{\sqrt{Var(W_+)}}$. W_+ is the sum of the ranks that have positive signs. Ties are assigned the average rank of the tied group. Since there are 400 test sentences, we have 400 paired samples where the elements of each pair are the AERs of the models being compared. The difference between Och and Tags at 5K, 10K, and 15K is significant at the $\sigma = 0.05$ level according to both tests. The difference between Och and Tags+Null is significant for all training set sizes at the $\sigma = 0.05$ level.

We also assessed the gains from using part of speech tags in the alignment probabilities according to the model described in section 5.2. Table 3 shows the error rate of the basic HMM alignment model as compared to an HMM model that conditions on tag sequences of source and target word tags in the neighborhood of the French word f_j and the English word $e_{a_{j-1}}$ for a training set size of 10K. The results we achieved showed an improvement of our model over a model that does not include conditioning on tags. The improvement in accuracy is best when using the current and previous French word parts of speech and does not increase when adding more conditioning information. The improvement from part of speech tag sequences for alignment probabilities was not as good as we had expected, however, which leads us to believe that more sophisti-

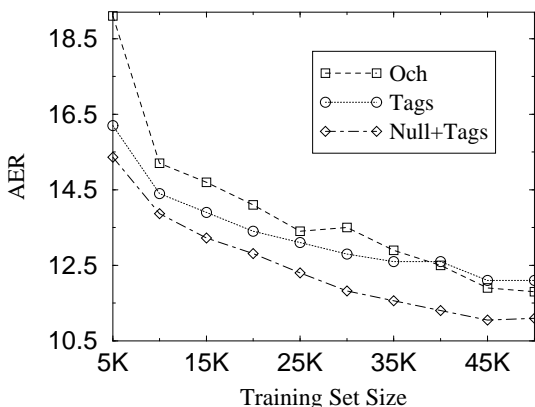


Figure 2: Och vs. Tags and Tags+Null.

Table 3: POS Conditioning of Jump Probabilities

Model	AER
Baseline	16.37
fT_j	15.97
$fT_j + fT_{j-1}$	15.74
$fT_j + eT_{a_{j-1}}$	15.86
$fT_j + eT_{a_{j-1}} + eT_{a_{j-1}+1}$	15.88
$fT_j + fT_{j-1} + eT_{a_{j-1}}$	15.94

cated syntax is needed to model local word order variation.

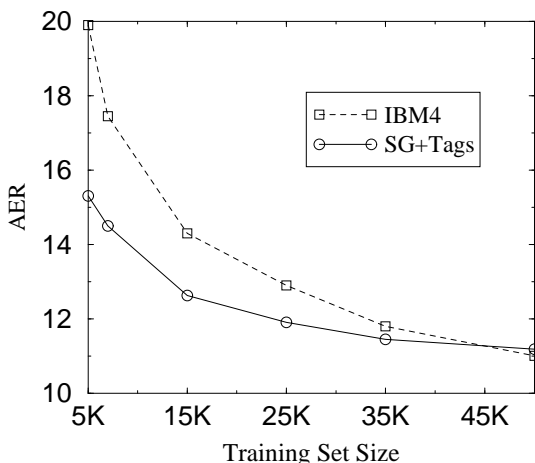


Figure 3: IBM-4 vs SG+Tags

In Figure 3 we compare the IBM-4 model to our SG+Tags model. Such a comparison makes sense because IBM-4 uses a fertility model for English words and SG approximates fertility modeling and because IBM-4 uses word classes as does our Tags model. For smaller training set sizes our model performs much better than IBM-4 but when more data is available IBM-4 becomes slightly better. This confirms the observation from Figure 2 that automatically induced bilingual classes perform better when trained on large amounts of data. Also as our fertility model estimates one parameter for each English word and IBM-4 estimates as many parameters as the maximum fertility allowed, at small training set sizes our model parameters can be estimated more reliably.

7 Conclusions

In this paper we presented three extensions to HMM-based alignment models. We showed that incorporating part of speech tag information of the source and target languages in the translation model improves word alignment accuracy. We also presented a method for approximately modeling fertility in an HMM-based model and a new generative model for target language words that do not have correspondences in the source language. The proposed models do not increase significantly the complexity of the learning algorithms while providing a better account for some phenomena in natural language translation.

References

- P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, volume 19(2), pages 263–311.
- Pascale Fung and Dekai Wu. 1995. Coerced markov models for cross-lingual tag relations. In *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, volume 1, pages 240–255.
- Dan I. Melamed. 2000. Models of translational equivalence among words. In *Computational Linguistics*, volume 26(2), pages 221–249.
- F. Och and H. Ney. 2000a. A comparison of alignment models for statistical machine translation. In *Proc. COLING '00: The 18th Int. Conf. on Computational Linguistics*, pages 1086–1090.
- F. Josef Och and H. Ney. 2000b. Improved statistical alignment models. In *Proc. of the 39th Annual Meeting of the ACL*.
- F. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine translation. In *Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.
- S. Vogel, H. Ney, and C. Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proc. COLING '96: The 16th Int. Conf. on Computational Linguistics*, pages 836–841.
- K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proc. of the 39th Annual Meeting of the ACL*, pages 523–530.