A Bayesian model for joint word alignment and part-of-speech transfer

Robert Östling
Department of Modern Languages, University of Helsinki
Department of Linguistics, Stockholm University
robert.ostling@helsinki.fi, robert@ling.su.se

Abstract

Current methods for word alignment require considerable amounts of parallel text to deliver accurate results, a requirement which is met only for a small minority of the world's approximately 7,000 languages. We show that by jointly performing word alignment and annotation transfer in a novel Bayesian model, alignment accuracy can be improved for language pairs where annotations are available for only one of the languages—a finding which could facilitate the study and processing of a vast number of low-resource languages. We also present an evaluation where our method is used to perform single-source and multi-source part-of-speech transfer with 22 translations of the same text in four different languages. This allows us to quantify the considerable variation in accuracy depending on the specific source text(s) used, even with different translations into the same language.

1 Introduction

Word alignment is the problem of identifying translationally equivalent words across the languages of a parallel text. It has found widespread use for enabling applications such as statistical machine translation (Brown et al., 1993; Koehn et al., 2003), annotation transfer (Yarowsky et al., 2001), word sense disambiguation (Diab and Resnik, 2002) and lexicon extraction (Wu and Xia, 1994).

Although many types of algorithms have been explored, the main line of research through the last couple of decades has been based on the generative IBM models introduced by Brown et al. (1993). What these models have in common is that they are unsupervised, asymmetric models, assuming one of the languages in a bitext (the *source language*) generates the corresponding text in the other language (the *target language*), word by word.

Most often, a variant of the Expectation-Maximization algorithm (Dempster et al., 1977) has been used for inference in these models, but recently there has been some work using Bayesian alignment models using Gibbs sampling for inference (DeNero et al., 2008; Mermer and Saraçlar, 2011; Gal and Blunsom, 2013). The incorporation of Bayesian priors into these models has been shown to improve accuracy, since they provide a flexible way of biasing the model towards empirical observations about language, most importantly that a given word type tends to have a very limited number of translations.

While the basic word alignment models use only lexical co-occurrence and word order, lexical data tends to be sparse and a number of authors have explored the usefulness of other information sources. Toutanova et al. (2002) showed that Part of Speech (PoS) tags can be integrated into the IBM models to improve word alignment accuracy, and others have reported similar results for dependency (Cherry and Lin, 2003; Wang and Zong, 2013) and phrase-structure (Yamada and Knight, 2001) parse trees, and for lemmatized texts (Bojar and Prokopov, 2006).

In addition to the studies just mentioned that showed how various types of linguistic annotation can be used to guide word alignment, there has been research showing that the reverse also holds: word-aligned parallel texts can be used to transfer annotations and models from languages where those resources exist to languages where they do not. Pioneering work by Yarowsky et al. (2001) explored tasks such as PoS tagging, shallow parsing and lemmatization, which was followed by e.g. dependency parsing (Hwa et al., 2005).

The present work combines these previous lines of work by exploring joint models of word alignment and annotation transfer (of PoS tags), within a Bayesian framework. The source code of our implementation is available at `http://www.ling.su.se/spacos`.

2 Methods

This section first discusses Bayesian word alignment using IBM-based models along with extensions to these, and finally describes our model of joint PoS transfer and word alignment.

2.1 Bayesian IBM models

The IBM 1 alignment model can be extended with sparse Dirichlet priors, and efficient inference is possible using Gibbs sampling (Mermer and Saraçlar, 2011; Mermer et al., 2013; Gal and Blunsom, 2013) or Variational Bayesian techniques (Riley and Gildea, 2012).

IBM model 1 assumes each target word $t_j = f$ of a sentence is generated by one source word $s_{a_j} = e$ through the alignment variable a_j, and that all words are generated independently and do not depend on the sentence positions i and j. The probability of a target sentence $\boldsymbol{t}$ (of length J) and an alignment $\boldsymbol{a}$ given a source sentence $\boldsymbol{s}$ (of length I) then becomes

$$P(\boldsymbol{t}, \boldsymbol{a}|\boldsymbol{s}) \propto p(J|I) \prod_{j=1}^{J} P(t_j|s_{a_j}) \tag{1}$$

One drawback of this model (apart from the extreme independence assumptions addressed by later IBM models) is that there is no penalty for having very flat distributions $P(f|e)$ for target words conditioned on a source word, a fact that causes the so-called *garbage collection effect* where rare source words are incorrectly linked to a large number of target words. By using priors on the translation distributions that discourage such solutions, it is possible to improve alignment accuracy. Mermer and Saraçlar (2011) introduced the use of Dirichlet priors for this task. If the Dirichlet parameter α is 1, this reduces to the uniform distribution, but it turns out that by using much smaller values of α, below about 10^{-2} (Riley and Gildea, 2012, Figure 1), the model better reflects the empirical observation that words tend to have very few possible translations.

2.2 Inference

For the standard IBM models, the EM algorithm is normally used for inference. In the Bayesian version with Dirichlet priors, we mentioned above that two main options have been investigated: Variational Bayes and Gibbs sampling. While both methods have been shown to improve word alignment accuracy for IBM model 1, the computational complexity of Gibbs sampling is lower with more complex models (Östling and Tiedemann, 2016, Section 3.2). For this reason, Gibbs sampling is used in the present work and will be discussed in further detail.

Gibbs sampling (Gelfand and Smith, 1991) is a specific instance of the more general Markov Chain Monte Carlo algorithm, which is used to draw samples from a model M which defines a probability function $p_M(\boldsymbol{x})$ over the variable space $\boldsymbol{x}$. This is done by constructing a Markov chain with p_M as its stationary distribution and performing a sufficiently long random walk in it. A Gibbs sampler achieves this by specifying for each variable x_i in $\boldsymbol{x}$ a sampling distribution $P(x_i = a|\boldsymbol{x}_{-i})$ for x_i conditioned on $\boldsymbol{x}_{-i}$, which denotes all variables in $\boldsymbol{x}$ except x_i. For IBM model 1, this gives the following sampling equation, which we also use, and for more complex models extend with additional factors given Equation (4):

$$P(a_j = i) = \frac{n_{\boldsymbol{a}_{-j}, s_i, t_j} + \alpha_{t_j}}{\sum_f (n_{\boldsymbol{a}_{-j}, s_i, f} + \alpha_f)} \tag{2}$$

Here, $n_{\boldsymbol{a}_{-j}, e, f}$ is a count vector representing the number of times each source type e is aligned to each target type f under the alignment $\boldsymbol{a}$, not counting the alignment at position j. In the end, we are interested in computing the expectations $\mathbb{E}\left[\delta_{a_j,i}\right]$ under the alignment model, where δ is the Kronecker delta. Given a series of samples of $\boldsymbol{a}^{(t)}$ for $t \in 1 \ldots T$, we approximate this using

$$\mathbb{E}\left[\delta_{a_j,i}\right] \approx \frac{1}{T}\sum_{t=1}^{T} P(a_j = i | \boldsymbol{a}_{-j}^{(t)}, \boldsymbol{s}, \boldsymbol{t}) \tag{3}$$

The initial alignments $\boldsymbol{a}^{(0)}$ are sampled from a uniform distribution, and in order to reduce initialization bias we average the marginals from eight independently initialized samplers. For details on the tradeoffs involved in choosing the number of samplers and sampling iterations, we refer to Table 2 of Östling and Tiedemann (2016).

2.3 Word order and fertility

Even with good prior parameters, IBM model 1 is a poor model of word alignment because it ignores two important characteristics of parallel text: word order and morpheme counts. While different distortion models have been used to model word order, we use the HMM-based model of Vogel et al. (1996), which has been demonstrated to deliver better performance than either no distortion model (like IBM model 1) or models based on absolute sentence positions (IBM models 2 and 3). This introduces a distribution $P(a_j - a_{j-1} | I)$ of the "jump" $a_j - a_{j-1}$ in the source sentence when moving one step in the target sentence.

As a way of modeling the relative number of morphemes in a word for a pair of languages, the *fertility* of a source word e is defined as the number of target words it is aligned to in a particular context. This is modeled using a distribution $P(\phi_i = k | s_i = e)$, where the fertility ϕ_i at position i is conditioned on the word e at that position. This is particularly important when the languages have large differences in word formation strategies and the general level of morphological complexity.

Zhao and Gildea (2010) explored a model with a word order and fertility model as described above, but based their work on the EM algorithm, using Gibbs sampling only for approximating the expectations. An important conclusion from their work is that a simple HMM with fertility model is competitive with the more complex IBM model 4, and we follow them in using this model as our baseline. Our full baseline model is given by

$$\begin{aligned} &P\big(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}\big) \\ &\quad \propto \left(\prod_{k=1}^{K}\prod_{j=1}^{J^{(k)}} \theta_{s^{(k)}_{a_j^{(k)}}, t_j^{(k)}}\right) \cdot \left(\prod_{e=1}^{E}\prod_{f=1}^{F} \theta_{e,f}^{\alpha_f - 1}\right) \\ &\quad \cdot \left(\prod_{k=1}^{K}\prod_{j=1}^{J^{(k)}+1} \psi_{a_j^{(k)} - a_{j-1}^{(k)}}\right) \cdot \left(\prod_{I=I_{min}}^{I_{max}}\prod_{m=m_{min}}^{m_{max}} \psi_m^{\beta_{I,m} - 1}\right) \\ &\quad \cdot \left(\prod_{k=1}^{K}\prod_{i=1}^{I^{(k)}} \pi_{s_i^{(k)}, \phi_i}\right) \cdot \left(\prod_{e=1}^{E}\prod_{n=0}^{n_{max}} \pi_{e,n}^{\gamma_n - 1}\right) \end{aligned} \tag{4}$$

where K is the number of parallel sentences, $\boldsymbol{\theta} \sim \text{Dir}\big(\boldsymbol{\alpha}\big)$ are the lexical translation parameters, $\boldsymbol{\psi} \sim \text{Dir}\big(\boldsymbol{\beta}\big)$ are the categorical distribution parameters for the word order model $P\big(a_j - a_{j-1} = m | I\big)$, and $\boldsymbol{\pi}_e \sim \text{Dir}\big(\boldsymbol{\gamma}\big)$ for the fertility model $P(\phi_i = k | s_i = e)$.

2.4 PoS-guided word alignment

This work follows Toutanova et al. (2002) in adding another factor to the model, akin to the lexical translation probability $P(f|e)$ but using the PoS tags of the respective words, $P(\boldsymbol{T}_f^t | \boldsymbol{T}_e^s)$. The main difference is that in their work PoS tags for both source and target languages were assumed, whereas here only one of the languages is assumed to be PoS-annotated. For the other language, the PoS tags are sampled using the method described below.

Algorithm 1 Alternating alignment-annotation.

```
▷ Align a single sentence pair $\boldsymbol{s}, \boldsymbol{t}$. The extension to multiple sentences is straightforward.
function AAA($\boldsymbol{s}, \boldsymbol{t}$)
    ▷ initialize alignments using the baseline HMM + fertility model
    ▷ the forward direction uses alignment vector $\boldsymbol{a}$
    $\boldsymbol{a} \leftarrow$ Baseline($\boldsymbol{s}, \boldsymbol{t}$)
    ▷ the backward direction uses alignment vector $\boldsymbol{b}$
    $\boldsymbol{b} \leftarrow$ Baseline($\boldsymbol{t}, \boldsymbol{s}$)
    while sampling do
        $M \leftarrow$ estimate bigram HMM model using $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{T}^s$
        ▷ set the target sentence tags $\boldsymbol{T}^t$ using the Viterbi algorithm
        $\boldsymbol{T}^t \leftarrow \arg\max_T P(T|M)$
        ▷ sample alignment variables $\boldsymbol{a}$, $\boldsymbol{b}$
        for all $j \leftarrow 1 \ldots J$ do
            $a_j \sim P(a_j = i | \boldsymbol{a}_{-j}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{T}^s, \boldsymbol{T}^t)$
        end for
        for all $i \leftarrow 1 \ldots I$ do
            $b_i \sim P(b_i = j | \boldsymbol{b}_{-i}, \boldsymbol{s}, \boldsymbol{t}, \boldsymbol{T}^s, \boldsymbol{T}^t)$
        end for
    end while
    ▷ return expected values for PoS tags alignments, as in Equation (3)
    return $\mathbb{E}[\boldsymbol{a}], \mathbb{E}[\boldsymbol{b}], \mathbb{E}[\boldsymbol{T}^t]$
end function
```

2.5 PoS transfer

We focus on applying PoS transfer as a way of obtaining better word alignment accuracy, rather than improving PoS transfer as such. These are largely complementary goals, as our evaluation in Section 3 shows that small changes in PoS tagging accuracy do not seem to influence alignment accuracy. For this reason, and because of our focus on low-resource languages precludes using data-intensive approaches like that of Das and Petrov (2011), we choose the simple method of Yarowsky and Ngai (2001) as a starting point for the PoS transfer part of our model. The basis of this method is to use heuristics to estimate a robust first-order HMM tagger from (noisy) projected tags, and to re-tag the data using this tagger. Furthermore we extended the tagger using the affix-tree method of Schmid (1994) for rare words, in order to be able to handle morphologically complex languages better.

While it would have been preferable for reasons of theoretical elegance to use a simpler PoS transfer model, matching the alignment model, such attempts by Östling et al. (2015) resulted in very modest improvements for their sign language data set, and their model gave no improvement at all for our data sets.

2.6 Alternating alignment-annotation (AAA)

Algorithm 1 summarizes our method, which can be viewed as a modified Gibbs sampler of the latent alignment variables $\boldsymbol{a}$ and $\boldsymbol{b}$ (in the forward and backward alignment direction) as well as the target-side PoS tags $\boldsymbol{T}^t$. While the PoS transfer part is not stochastic,[1] it operates on samples of the alignment variables $\boldsymbol{a}$ and $\boldsymbol{b}$ and can be seamlessly integrated into the sampler.

3 Evaluation

The empirical evaluation aims at investigating whether the alternating alignment-annotation (AAA) algorithm improves word alignment and/or PoS transfer accuracy, compared to the corresponding PoS-

[1] We also tried sampling $\boldsymbol{T}^t$ using marginal distributions computed by the forward-backward algorithm, but found no effect on the accuracy of the algorithm.

Corpus	**Sentences**	$\lvert S\rvert$	$\lvert P\rvert$
English-French	1 130 588	4 038	17 438
Romanian-English	48 641	5 034	5 034
English-Inuktitut	333 185	293	1 972
English-Hindi	3 556	1 409	1 409
English-Swedish	692 662	3 340	4 577

Table 1: Total corpus sizes (in sentences) and number of (S)ure and (P)ossible alignment links in their respective evaluation sets.

unaware Bayesian IBM model with an HMM word order model and fertility.

3.1 Data

In order to assess the general usefulness of the method presented, a number of parallel corpora representing a diverse set of languages and genres are used: the English-French Hansards corpus in the version presented by Mihalcea and Pedersen (2003), the Romanian-English, English-Inuktitut and English-Hindi corpora from Martin et al. (2005), as well as parts of the Swedish-English Europarl corpus (Koehn, 2005) with the evaluation set of Holmqvist and Ahrenberg (2011). In addition, a set of translations of the New Testament will be used to investigate the quality of the transfered PoS tags. Some properties of these corpora are summarized in Table 1.

Silver-standard PoS annotations were provided for English, French and German by the Stanford Tagger (Toutanova et al., 2003) and for Swedish by Stagger (Östling, 2013). The native tagsets were mapped to the Universal PoS Tagset of Petrov et al. (2012).

3.2 Experimental setup

The main intended use case is a fairly short parallel text, with two very different languages of which only one has an accurate PoS tagger available. This excludes the possibility of extensive per-language tuning (unlike some of the previous results cited), and in this evaluation the different language pairs use identical parameters to the largest possible extent.[2] We fixed the hyperparameters in Equation (4) to $\alpha = 10^{-5}$, $\beta = \gamma = 1$.

The experiments use eight individually initialized samplers, each of which used a pipelined approach where initially a lexical-only model equivalent to that of Mermer and Saraçlar (2011) was used, then a word order term using the HMM model was added, then the fertility term, and finally (when applicable) the PoS translation probability. No burn-in period was used during sampling, since the initial value of the last pipeline step is already quite good.

Since the model is asymmetric, the alignments are run in both directions and symmetrized. A soft variant of the intersection heuristic is used, where the final set of links L is defined as $L = \{(i, j) \mid P(a_j = i)P(b_i = j) > t\}$. for a threshold value t, in these evaluations fixed to 0.25. This gives a fairly conservative set of links, favoring precision before recall. Note however that the model does not use NULL words, so this conservatism is not as severe as in models with NULL words.[3] Heuristics based on the union on the contrary tend to over-generate links under these conditions.

3.3 Results

The systems used as baselines in the evaluation are mainly from the Workshop on Parallel Text shared tasks (Mihalcea and Pedersen, 2003; Martin et al., 2005), where most of the data sets used were intro-

[2] The main exception is that some of the language pairs (Romanian-English and English-Hindi), following previous work, use a poor man's stemming trick where only the first four letters of each token is used. The only other exception is that the English-French evaluation did not use the fertility parameter, since it showed no further improvement beyond the plain HMM model.

[3] Later experiments with a related model (Östling and Tiedemann, 2016, compare their Table 2 with our Table 2) show that for some language pairs in particular, using NULL with standard symmetrization heuristics gives considerably worse AER scores.

duced: ISI2 (Fraser and Marcu, 2005), JHU (Schafer and Drábek, 2005), UMIACS2 (Lopez and Resnik, 2005) and XRCE (Dejean et al., 2003). The Swedish-English figures, LIU, are from Holmqvist and Ahrenberg (2011). Finally, we have run GIZA++ (Och and Ney, 2003) on the corpora as an additional baseline.[4] Results from previous work using more data than a bitext plus PoS tags for one of the languages are not included, although some such systems have obtained better results on some of the corpora used, using e.g. semi-supervised discriminative training (Liu et al., 2010).

Table 2 summarizes the main results of the evaluation. In all cases, the alternating alignment-annotation method surpasses the baseline model that does not use PoS tags. The model is generally competitive compared to previous work, in particular for the smaller corpora and where the languages are substantially different. The improvement compared to the non-Bayesian baselines is particularly good for the English-Inuktitut corpus, which could be due to the fact that the morpheme/word ratio of Inuktitut is very high, resulting in very many low-frequency words that tend to function as garbage collectors in non-Bayesian models. The situation is similar for the English-Hindi corpus, although in this case the cause for the many rare words is rather the short bitext than the languages themselves.

It is interesting to compare the corresponding **AAA** and **Supervised** figures in Table 2, where the only difference is that **AAA** uses a supervised PoS tag on one language (English) and annotation transfer to the other, whereas **Supervised** uses supervised PoS tags on both languages. The overall accuracy figures are nearly identical, even though the accuracy of the transfered tags is lower. This indicates that the word alignment algorithm is not very sensitive to PoS tagging accuracy, so that the relatively simple PoS transfer method used is sufficient for the purpose of increasing word alignment accuracy.

There are multiple translations of the New Testament into each of English, French, German and Swedish, which can be exploited for multi-source transfer. In our model, multi-source transfer can be done trivially by averaging the expectations returned by Algorithm 1. Table 3 shows that this overall has a large positive effect on PoS accuracy, with an average error reduction of one fourth compared to the median single-source result, and one tenth compared to the best (out of 22) single-source result. Using many translations in each language allows us to see how widely the accuracy varies, even when using the same source (or target) language. This is due to many factors, including the large time span (hundreds of years) between the different translations. In contrast, the multi-source results are, as could be expected, much more robust. This is an encouraging result, given that the New Testament is perhaps the most widely translated text of significant length, and offers a great possibility to transfer linguistic annotations to languages where little other data is available.

4 Conclusions and future work

We have presented a model for joint word alignment and PoS annotation transfer, and shown empirically that it leads to improved word alignment accuracy, in particular for low-resource languages. Using automatically transfered PoS tags led to improvements that were as big as the improvements seen when using PoS tags from supervised taggers on both sides of a bitext.

In addition, we took the opportunity to perform an evaluation investigating what kind of variation can be expected depending on which translation(s) are used as source texts in PoS annotation transfer, and found that this variation can be great, even among translations into the same language. Using multi-source transfer reduces this variation considerably and typically gives better accuracy than even the best single-source transfer among many.

In this study, only PoS annotations were considered, but there are other types of annotation such as parse trees, named entities and word senses which potentially could be transfered jointly with word alignment. This is left to future work, as are improvements to the baseline alignment model.

Acknowledgments

Thanks to the anonymous reviewers, Mats Wirén, Jörg Tiedemann and Joakim Nivre for advice.

[4] In order to provide a competitive but fair baseline, the same general approach was used with GIZA++ as with the new system presented, using default parameters and no language-specific tuning. The specific alignment pipeline used was $1^3 h^5 3^3 4^{10}$, and the symmetrization that provides the best alignment on the test set is chosen. This gives GIZA++ some advantage, but ensures that any claimed improvements by our algorithm over GIZA++ are not simply due to symmetrization.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
English-French ($\|S\| = 4038$, $\|P\| = 17438$. 1 130 588 sentences)							
Baseline	5359	3717	5134	95.8	92.1	93.9	5.8
AAA	5505	3751	5254	95.4	92.9	94.1	5.6
Supervised	5542	3778	5263	95.0	93.6	94.3	5.6
GIZA++	4831	3531	4715	97.6	87.4	92.2	7.0
XRCE				90.1	93.8	91.9	8.5
Romanian-English ($\|S\| = \|P\| = 6201$. 48 641 sentences)							
Baseline	3374	3070	3070	91.0	61.0	73.0	27.0
AAA	3447	3120	3120	90.5	62.0	73.6	26.4
GIZA++	3730	3161	3161	84.7	62.8	72.1	27.9
ISI2				87.9	63.1	73.5	26.6
RACAI				76.8	71.2	73.9	26.1
English-Inuktitut ($\|S\| = 293$, $\|P\| = 1972$. 333 185 sentences)							
Baseline	598	267	559	93.5	91.1	92.3	7.3
AAA	630	273	595	94.4	93.2	93.8	6.0
GIZA++	342	170	306	89.5	58.0	70.4	25.0
JHU				96.7	76.8	85.6	9.5
JHU				84.4	92.2	88.1	14.3
English-Hindi ($\|S\| = \|P\| = 1409$. 3 556 sentences)							
Baseline	712	606	606	85.1	43.0	57.1	42.9
AAA	817	677	677	82.9	48.0	60.8	39.2
GIZA++	984	615	615	62.5	43.6	51.4	48.6
UMIACS2				43.7	56.1	49.1	50.9
English-Swedish ($\|S\| = 3340$, $\|P\| = 4577$. 692 662 sentences)							
Baseline	3183	2742	2933	92.1	82.1	86.8	13.0
AAA	3125	2774	2961	94.8	83.1	88.5	11.3
Supervised	3262	2823	3034	93.0	84.5	88.6	11.3
GIZA++	3436	2890	3136	91.3	86.5	88.8	11.1
LIU				85.3	–	–	12.6

Table 2: Results from the empirical evaluation, including the Bayesian model without PoS tags (Baseline), the alternating alignment-annotation algorithm (AAA), the corresponding method but with supervised PoS taggers for both languages (Supervised), and comparable previous results on the same data. The number of alignment links $|A|$, of which $|A \cap S|$ are considered (S)ure, and $|A \cap P|$ (P)ossible, are reported. For convenience, precision (P), recall (R), F_1 score (F) and Alignment Error Rate (AER) are also given.

Target	Source texts															Multi
	deu (8)			eng (5)			fra (5)			swe (4)			All (22)			
deu$_1$				79.0	80.1	80.9	80.1	81.1	81.4	75.0	83.7	85.1	75.0	81.0	85.1	**86.8**
deu$_2$				79.3	80.8	81.4	79.5	80.5	80.7	78.3	83.5	85.2	78.3	80.7	85.2	**85.8**
deu$_3$				79.8	80.6	81.7	81.1	81.9	82.0	77.1	84.4	85.9	77.1	81.7	85.9	**88.2**
deu$_4$				80.6	81.1	82.4	81.0	81.8	81.8	75.3	83.2	85.4	75.3	81.6	85.4	**87.3**
deu$_5$				80.2	80.7	81.7	81.3	81.6	82.2	76.5	84.9	85.9	76.5	81.5	85.9	**86.8**
deu$_6$				79.4	81.3	81.9	80.0	80.7	81.3	80.7	85.0	**86.0**	79.4	81.0	**86.0**	85.4
deu$_7$				79.9	81.8	82.3	81.1	81.2	81.9	76.6	85.6	86.3	76.6	81.7	86.3	**86.4**
deu$_8$				80.0	81.4	82.3	81.0	81.4	82.0	76.3	84.8	85.7	76.3	81.6	85.7	**86.5**
eng$_1$	74.2	76.2	79.4				76.2	77.0	77.8	76.6	81.5	81.7	74.2	76.7	81.7	**83.8**
eng$_2$	79.2	81.8	84.2				80.9	81.4	82.0	79.8	85.8	**86.2**	79.2	81.8	**86.2**	85.5
eng$_3$	80.1	81.7	83.7				80.6	81.0	81.6	80.7	85.7	**86.3**	80.1	81.6	**86.3**	85.4
eng$_4$	79.5	81.4	84.3				80.3	80.7	81.3	78.8	85.8	**86.5**	78.8	81.3	**86.5**	85.1
eng$_5$	80.3	81.5	84.1				80.7	81.0	81.8	80.3	86.0	**86.7**	80.3	81.5	**86.7**	84.7
fra$_1$	80.2	82.9	83.5	80.1	81.3	81.5				78.0	83.8	84.5	78.0	82.5	84.5	**85.8**
fra$_2$	80.3	83.5	84.5	80.7	80.9	81.1				77.0	83.6	84.7	77.0	83.0	84.7	**86.0**
fra$_3$	80.5	83.1	83.5	79.9	81.2	81.9				77.6	84.3	85.1	77.6	82.7	85.1	**85.3**
fra$_4$	80.3	83.5	83.8	80.0	81.1	81.2				77.4	84.1	84.6	77.4	82.7	84.6	**85.3**
fra$_5$	80.5	83.2	83.8	80.1	81.2	81.4				77.2	84.0	84.8	77.2	82.9	84.8	**86.1**
swe$_1$	80.4	81.2	81.8	82.8	84.0	85.4	80.2	81.0	81.9				80.2	81.2	85.4	**90.3**
swe$_2$	76.3	77.5	79.4	80.9	81.7	82.4	76.9	77.9	79.4				76.3	78.3	82.4	**85.7**
swe$_3$	81.3	82.1	82.5	83.4	85.4	86.4	81.1	82.5	82.8				81.1	82.5	86.4	**90.6**
swe$_4$	81.8	82.3	82.7	82.7	84.8	85.4	81.8	82.7	83.3				81.8	82.7	85.4	**90.7**
Avg.	79.6	81.6	82.9	80.5	81.7	82.4	80.2	80.9	81.5	77.7	84.4	85.4	77.9	81.5	85.3	**86.5**

Table 3: PoS transfer accuracy (in percent) using single-source (first five columns) and multi-source (rightmost column) transfer in the New Testament corpus. Rows are target texts and columns are source languages. For each language (with number of translations), the worst/median/best results are given for the different translations. The **All** columns summarize the results over all the source texts from the preceding columns. Finally, **Multi** is the result of multi-source transfer using the sums of tag marginal distributions. The best result on each row is bold-faced.

References

Ondej Bojar and Magdalena Prokopov. 2006. Czech-English word alignment. In *LREC 2006*, pages 1236–1239, Genova, Italy. ELRA.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *ACL 2003*, pages 88–95, Sapporo, Japan, July. Association for Computational Linguistics.

Dipanjan Das and Slav Petrov. 2011. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *ACL 2011*, HLT '11, pages 600–609, Stroudsburg, PA, USA. Association for Computational Linguistics.

Herve Dejean, Eric Gaussier, Cyril Goutte, and Kenji Yamada. 2003. Reducing parameter space for word alignment. In *HLT-NAACL-PARALLEL '03*, pages 23–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

John DeNero, Alexandre Bouchard-Côté, and Dan Klein. 2008. Sampling alignment structure under a Bayesian translation model. In *EMNLP 2008*, pages 314–323, Honolulu, Hawaii, October. Association for Computational Linguistics.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *ACL 2002*, pages 255–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2005. ISI's participation in the Romanian-English alignment task. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 91–94, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yarin Gal and Phil Blunsom. 2013. A systematic Bayesian treatment of the IBM alignment models. In *NAACL 2013*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alan E. Gelfand and Adrian F. M. Smith. 1991. Gibbs sampling for marginal posterior expectations. Technical report, Department of Statistics, Stanford University.

Maria Holmqvist and Lars Ahrenberg. 2011. A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, number 11 in NEALT Proceedings Series, pages 106–113.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit.*, Phuket, Thailand.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339, September.

Adam Lopez and Philip Resnik. 2005. Improved HMM alignment models for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 83–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 65–74, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coşkun Mermer and Murat Saraçlar. 2011. Bayesian word alignment for statistical machine translation. In *ACL 2011*, pages 182–187, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coşkun Mermer, Murat Saraçlar, and Ruhi Sarikaya. 2013. Improving statistical machine translation using Bayesian word alignment and Gibbs sampling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):1090–1101, May.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with Markov Chain Monte Carlo. *Prague Bulletin of Mathematical Linguistics*, 106:125–146, October.

Robert Östling, Carl Börstell, and Lars Wallin. 2015. Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015)*, volume 23 of *NEALT Proceedings Series*, pages 263–268, Vilnius, Lithuania, May.

Robert Östling. 2013. Stagger: An open-source part of speech tagger for Swedish. *North European Journal of Language Technology*, 3:1–18.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *LREC 2012*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Darcey Riley and Daniel Gildea. 2012. Improving the IBM alignment models using variational Bayes. In *ACL 2012*, pages 306–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

Charles Schafer and Elliott Franco Drábek. 2005. Models for Inuktitut-English word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05, pages 79–82, Stroudsburg, PA, USA. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Kristina Toutanova, H. Tolga Ilhan, and Christopher Manning. 2002. Extensions to HMM-based statistical word alignment models. In *EMNLP 2002*, pages 87–94.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL 2003*, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING 1996*, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhiguo Wang and Chengqing Zong. 2013. Large-scale word alignment using soft dependency cohesion constraints. *Transactions of the Association for Computational Linguistics*, 1:291–300.

Dekai Wu and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 206–213.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 523–530, Toulouse, France, July. Association for Computational Linguistics.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *NAACL 2001*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *HLT 2001*, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shaojun Zhao and Daniel Gildea. 2010. A fast fertility Hidden Markov Model for word alignment using MCMC. In *EMNLP 2010*, pages 596–605, Cambridge, MA, USA, October. Association for Computational Linguistics.