

Translation Lexicon Estimates from Non-Parallel Corpora Pairs

Markos Mylonakis

Khalil Sima'an

*Language and Computation, University of Amsterdam, Plantage Muidersgracht 24,
1018TV Amsterdam. {mmylonak, simaan}@science.uva.nl.*

Abstract

The estimation of translation lexicon probabilities from parallel corpora is well studied in statistical machine translation. Whenever parallel corpora are not available, it is still possible to obtain unsupervised estimates from pairs of monolingual, non-parallel corpora. In both cases the standard estimator is the Expectation-Maximization (EM) that aims at increasing the likelihood of the source corpus under the translation model. In this paper we study the utility of maximizing the joint likelihood of source and target corpora under a bi-directional translation model. A recently presented bi-directional estimator (Bi-EM [10]), which maximizes the joint likelihood, constitutes an instance of the EM algorithm. We show that Bi-EM reconciles the asymmetric statistics in the two corpora and leads to better lexicon estimates than standard EM. Our extensive experimental results show that relative to standard EM, our Bi-EM gives substantially better word-to-word translation results across variably related pairs of monolingual corpora.

1 Introduction

Word translation probabilities can be useful for improving statistical alignment and machine translation models, e.g. [2, 8, 14], and for leveraging language processing resources from a resource-rich to a resource-poor language, e.g. [13, 3]. The conditional probability $p(e|f)$ expresses the prior probability of a target word e being the correct translation of source word f regardless of context.

Whenever a large parallel corpus is available, such probabilities can be acquired by relative frequency estimates over automatic word-to-word alignment, e.g. [8]. Unfortunately, sufficiently large parallel corpora are not always available. In particular, for some language pairs, e.g. a language and its spoken (but hardly ever written) dialect, one cannot expect to find a parallel corpus. The alternative is to estimate translation probabilities for the entries of a *translation lexicon* over pairs of monolingual (non-parallel, not necessarily related) corpora. Such translation lexica can be built manually (dictionaries) or acquired semi-automatically.

In this paper we consider the problem of *unsupervised* estimation of translation probabilities for the entries of a translation lexicon from pairs of non-parallel corpora. We start out from the simple noisy-channel translation model described by [7] and explore the utility of various unsupervised estimators. A straightforward choice for an unsupervised estimator is the Expectation-Maximization algorithm [5, 1] explored by [7]. This estimator aims at adjusting the lexicon probabilities in order to increase the likelihood of the source corpus under the given translation model.

Underlying any translation task lies the assumption that the translation lexicon constitutes a mapping between two corpora, *regardless of translation direction*. Based on this observation, we explore a new estimator that aims at meeting the statistical constraints posed by *both* translation directions, source \rightleftarrows target. This estimator (see also [10]) aims at maximizing the *joint likelihood of the source and target monolingual corpora*. We show how this Maximum-Likelihood estimator can be implemented as a bi-directional EM (Bi-EM) algorithm that provides estimates that fit both corpora better than estimates from the source side only.

Interestingly, [14] present a related algorithm for estimating lexicon probabilities from *parallel corpora* in order to improve word alignment. While their aim is largely the same as ours, Zens et al. arrive at their algorithm as a *crude approximation* for the solution of a complex maximization involving an interpolation over the expectations obtained from the two translation directions. The algorithm turns out neither an EM

instance nor does it clear what objective function of the data it is optimizing. Furthermore, for the approximation made by Zens et al. to be reasonable, a strong assumption is needed: the unigram counts in both corpora must remain unchanged regardless of translation directions. Here we show that when the estimator maximizes the joint likelihood there is no need for such assumptions. Furthermore, we arrive at the Bi-EM algorithm from the well-understood Maximum-Likelihood approach rather than as an interpolation at the algorithmic level.

We apply the present algorithms to estimate the probabilities for a given lexicon from pairs of non-parallel corpora of varying degrees of relatedness to one another (different domains). To facilitate automatic evaluation of the probability estimates, they are embedded in a word-to-word translation system that we apply to a standard translation task for which we have a gold-standard parallel corpus. Our experiments show that the Bi-EM outperforms other existing methods significantly, specifically unidirectional-EM [7], even with less than half the training data and regardless of the level of relatedness of the monolingual corpora to one another.

This paper is structured as follows: Section 2 discusses related work and includes the model and baseline estimator on which we improve. Section 3 presents the joint likelihood estimator (JLE) and the Bi-EM algorithm. Section 3 reviews implementation detail. Section 4 presents extensive experiments that show the utility of the Bi-EM algorithm in a simple translation task. Finally section 5 gives the conclusions from this work.

2 Related work

Lexicon estimates in SMT For a source sentence $\mathbf{f} = (f_1, \dots, f_n)$ and a target sentence $\mathbf{e} = (e_1, \dots, e_m)$, statistical machine translation approaches often start out from the noisy channel [2]:

$$\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) = \max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e})p(\mathbf{e})$$

In a large majority of SMT work, a parallel corpus is employed (see e.g. [2, 8]) and it is assumed that a hidden alignment \mathbf{a} can be built between the words of each pair of aligned sentences \mathbf{f} and \mathbf{e}

$$\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) = \max_{\mathbf{e}} \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e})p(\mathbf{e})$$

For estimating the word alignment probabilities and the lexicon probabilities, most work employs the Expectation-Maximization (EM) algorithm [5], starting from an impoverished alignment model (IBM model 1) to obtain initial estimates of the translation lexicon probabilities. The initial lexicon estimates are then used in a more complex alignment algorithm to obtain alignment estimates. When the parallel corpus is aligned at the word-level, translation probabilities can be re-estimated by relative frequency [8] or again by EM.

Baseline model In contrast with work using parallel corpora, in [7] as well as in the present case, only a pair of monolingual corpora is available. Clearly, there is no possible alignment between the pairs of sentences, instead an ambiguous translation lexicon L is assumed provided. For every word f , L contains a set of translations $L(f)$, and vice versa (for e it contains a set $L(e)$). The goal is to estimate translation probabilities $p(f|e)$, the probability that a word e translates as word $f \in L(e)$, regardless of context. Let the set $L(\mathbf{f})$ stand for the set (or lattice) of all possible target sentences \mathbf{e} that result from translating the (ordered) sequence of words in \mathbf{f} , one by one¹, using lexicon L . Koehn and Knight derive the following model:

$$\max_{\mathbf{e} \in L(\mathbf{f})} p(\mathbf{e}|\mathbf{f}) = \max_{\mathbf{e} \in L(\mathbf{f})} p_{\overleftarrow{\theta}}(\mathbf{f}|\mathbf{e})p(\mathbf{e}) = \max_{\mathbf{e} \in L(\mathbf{f})} p(\mathbf{e}) \prod_{i=1}^n \overleftarrow{\theta}(f_i|e_i) \quad (1)$$

where $\overleftarrow{\theta}$ stands for the translation lexicon probabilities $\mathbf{f} \leftarrow \mathbf{e}$, i.e. $p(\mathbf{f}|\mathbf{e})$. This model employs a language model $p(\mathbf{e})$ over target sentences trained on the target language monolingual corpus \mathcal{E} , and a “translation model” with lexicon probabilities $\overleftarrow{\theta}(f_i|e_i)$.

¹Thereby assuming the same word-order and a one-to-one mapping between words, which also implies that sentence length is unchanged, i.e. $m == n$.

Using fixed language model parameters $\tilde{p}(\mathbf{e})$, the lexicon probabilities are estimated using the Expectation Maximization (EM) algorithm [5] over the source language corpus \mathcal{F} . Assuming an initial estimate $\overleftarrow{\theta}_0$ for $\overleftarrow{\theta}$, and denote the current estimate at iteration r by $\overleftarrow{\theta}_r$.

E-step_r: for every $\mathbf{f} \in \mathcal{F}$ and $\mathbf{e} \in L(\mathbf{f})$: $q_r(\mathbf{e}|\mathbf{f}) := \frac{1}{Z_r(\mathbf{f})} \tilde{p}(\mathbf{e}) \prod_{i=1}^n \overleftarrow{\theta}_r(f_i|e_i)$

M-step_r: maximize over $\overleftarrow{\theta}$ to obtain $\overleftarrow{\theta}_{r+1} := \arg \max_{\overleftarrow{\theta}} \sum_{\substack{\mathbf{f} \in \mathcal{F} \\ \mathbf{e} \in L(\mathbf{f})}} q_r(\mathbf{e}|\mathbf{f}) \log[\tilde{p}(\mathbf{e}) p_{\overleftarrow{\theta}}(\mathbf{f}|\mathbf{e})]$

Where $Z_r(\mathbf{f}) = \sum_{\mathbf{e} \in L(\mathbf{f})} \tilde{p}(\mathbf{e}) \prod_{i=1}^n \overleftarrow{\theta}_r(f_i|e_i)$. The maximization at iteration r (M-step_r) is calculated by relative frequency estimates as follows:

$$\overleftarrow{\theta}_r(f|e) = \frac{\sum_{\substack{\mathbf{f} \in \mathcal{F} \\ \mathbf{e} \in L(\mathbf{f})}} q_r(\mathbf{e}|\mathbf{f}) \times \sum_j \delta[f_j, f] \delta[e_j, e]}{\sum_{\substack{\mathbf{f} \in \mathcal{F} \\ \mathbf{e} \in L(\mathbf{f})}} q_r(\mathbf{e}|\mathbf{f}) \times \sum_j \delta[e_j, e]}$$

where $\delta[x, y] = 1$ iff $x = y$, and zero otherwise. The actual implementation for Hidden Markov Models is known as the Baum-Welch or Forward-Backward algorithm [1].

2.1 Existing bi-directional estimation methods

It has been observed in the SMT literature that intersecting the alignments estimated from the two possible directions of translation $\mathcal{F} \rightarrow \mathcal{E}$ and $\mathcal{F} \leftarrow \mathcal{E}$ improves the precision of the alignment [11]. However, intersecting alignments do not provide probability estimates. Reconciling the alignments of the two directions of translation culminates in the method of [14] for the estimation of lexicon probabilities from parallel corpora. This method employs two directional translation models, each with a hidden directional alignment model and a word-to-word lexicon. The crucial observation of Zens et al., shared with our approach, is that the conditional lexicon probabilities can be computed using joint estimates from counts over the alignments obtained from either translation direction. Contrary to our approach, however, Zens et al. employ *two separate* Uni-EM algorithms to construct two probabilistic directional alignments. After each iteration of these Uni-EM algorithms, each of the directional alignments is used for acquiring estimates of the joint counts for the lexicon word-pairs. These joint counts are then interpolated together leading to “symmetrized” lexicon probability estimates, which are in turn fed back into each of the separate Uni-EM algorithms. It is unclear what objective function of the data this method is optimizing. Furthermore, Zens et al. make unrealistic and unnecessary assumptions regarding the unigram counts in the two corpora.

Coming up to date, [9] present “Agreement Alignment”: The key idea is to employ the parallel corpus $\langle \mathcal{F}, \mathcal{E} \rangle$ for the estimation of two alignments $\overleftarrow{\theta}$ and $\overrightarrow{\theta}$ (the two directions of translation) under an objective likelihood function of $\langle \mathcal{F}, \mathcal{E} \rangle$ that measures individual fit to the data as well as mutual “agreement” between these alignments:

$$\mathbb{L}(\mathcal{F}, \mathcal{E}; \overrightarrow{\theta}) \times \mathbb{L}(\mathcal{F}, \mathcal{E}; \overleftarrow{\theta}) \times \mathbb{L}(\mathcal{F}, \mathcal{E}; Agr(\overrightarrow{\theta}, \overleftarrow{\theta}))$$

where $\mathbb{L}(X; \theta) = \prod_{x \in X} p_{\theta}(x)$ stands for the likelihood of parallel corpus X (sentence pairs) under the (translation) model that employs alignment θ , and $Agr(a, b)$ measures the agreement between the two alignments a and b given $x \in X$ as the dot product of two probability vectors that range over all possible alignments between that pair (also called set of generalized alignments).

While the idea of agreement alignment is appealing, it is by definition not directly applicable in the present case as we start out from a non-parallel corpus. Furthermore, because the lexicon is large (relative to sentence length), it is computationally and statistically prohibitive to employ the same measure of agreement (such as dot product) between the two estimates of probabilities (per direction) over the subsets of the translation lexicon (the power set of the lexicon). Apart from these technical objections, we present here an alternative, more appealing approach to conduct estimation under agreement constraints between two directions of translation.

3 Joint-Likelihood estimators

We review the principles that underly Bi-EM [10] before specifying its working. We depart from the intuition that the independent estimation of the lexicon probabilities $\tilde{p}_{\overleftarrow{\theta}}(\mathbf{f}|\mathbf{e})$ and $\tilde{p}_{\overrightarrow{\theta}}(\mathbf{e}|\mathbf{f})$ yields empirical estimates

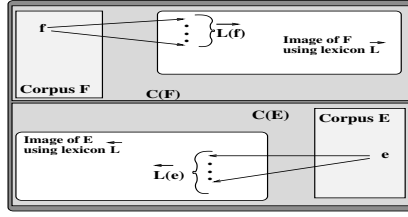


Figure 1: Schematic view of \mathcal{C}_T .

that *do not agree* on the joint probability $p(\mathbf{f}, \mathbf{e})$, i.e.

$$\tilde{p}(\mathbf{e})\tilde{p}_{\leftarrow}(\mathbf{f}|\mathbf{e}) \neq \tilde{p}(\mathbf{f})\tilde{p}_{\rightarrow}(\mathbf{e}|\mathbf{f})$$

This inequality is expected due to the “asymmetric” statistics in \mathcal{E} and \mathcal{F} and the way each is used in each (directed) model. We hypothesize that the notion of “agreement” between the two models can be implemented by estimation under the constraint that consensus is achieved over this joint probability. A naive approach would be to take the weighted sum of the final EM estimates obtained over the two translation directions (each conducted on its own):

$$\tilde{p}(\mathbf{f}, \mathbf{e}) = \lambda\tilde{p}_{\leftarrow}(\mathbf{f}, \mathbf{e}) + (1 - \lambda)\tilde{p}_{\rightarrow}(\mathbf{f}, \mathbf{e}) \quad (2)$$

where λ could be, e.g. the ratio of corpora sizes. This leads to re-estimates $p_{\leftarrow}(\mathbf{f}|\mathbf{e}) = \frac{\tilde{p}(\mathbf{f}, \mathbf{e})}{\sum_{\tilde{\mathbf{f}}} \tilde{p}(\tilde{\mathbf{f}}, \mathbf{e})}$ and $p_{\rightarrow}(\mathbf{e}|\mathbf{f}) = \frac{\tilde{p}(\mathbf{f}, \mathbf{e})}{\sum_{\tilde{\mathbf{e}}} \tilde{p}(\mathbf{f}, \tilde{\mathbf{e}})}$. The Average-EM re-estimates are obtained only *after training*, in analogy to the intersection of alignments in SMT. Despite of being attractive, it is unclear what objective function these reestimates aim at.

Our approach aims at maximizing the joint-likelihood of the two corpora under a joint probability model $p_{\theta}(\mathbf{f}, \mathbf{e}) = \prod_{i=1}^n \theta(f_i, e_i)$ which coordinates two internally hidden conditional, directed translation models that are both employing the same set of translation parameters θ . Let $p_1(\mathbf{f})$ be a language model estimated from \mathcal{F} and analogously $p_2(\mathbf{e})$ from \mathcal{E} , we rewrite the directional translation models in terms of a single set of lexicon parameters θ :

$$\max_{\mathbf{f}} p(\mathbf{f}|\mathbf{e}) = \max_{\mathbf{f}} p_1(\mathbf{f}) \frac{p_{\theta}(\mathbf{e}, \mathbf{f})}{\sum_{\tilde{\mathbf{e}}} p_{\theta}(\tilde{\mathbf{e}}, \mathbf{f})} \quad \max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) = \max_{\mathbf{e}} p_2(\mathbf{e}) \frac{p_{\theta}(\mathbf{e}, \mathbf{f})}{\sum_{\tilde{\mathbf{f}}} p_{\theta}(\mathbf{e}, \tilde{\mathbf{f}})}$$

Stating the two models in terms of the same set of joint probabilities of words implies that the source and target corpora are assumed to have been generated from a single source: the joint lexicon probabilities. This allows us to state a new objective function, the Joint-Likelihood of two monolingual corpora:

$$\begin{aligned} \max_{\theta} \mathbb{L}(\mathcal{E}; \theta, p_1, L) \times \mathbb{L}(\mathcal{F}; \theta, p_2, L) \\ \mathbb{L}(X; \theta, p_k, \hat{L}) = \prod_{\mathbf{x} \in X} \sum_{\mathbf{y} \in \hat{L}(\mathbf{x})} p_k(\mathbf{y}) \frac{p_{\theta}(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{x}'} p_{\theta}(\mathbf{x}', \mathbf{y})} \end{aligned} \quad (3)$$

This statement of the objective function optimizes over θ the joint-likelihood of two monolingual corpora, each under its own likelihood function which involves the other corpus. The joint-likelihood function is more restricted than the likelihood function of the source corpus alone because it involves both sides of the translation channel. Our hypothesis is that the optimization of the joint-likelihood is as meaningful as the unidirectional version because both make the assumption that the two corpora are connected by some statistical translation relation (expressed in the lexicon), which hinges on their relatedness in terms of genre, mode, style, domain and other corpus features. We will explore the effect of various pairs of monolingual training corpora on the estimation algorithms in section 4.

Crucially, the joint likelihood function has the same form as the usual likelihood function with the minor difference that the multiplication ranges over two rather than one corpus (each under its own translation direction). In the light of this observation we can directly obtain a Bidirectional-EM algorithm that aims at the joint-likelihood, just in the same fashion the EM is obtained from standard maximum-likelihood.

Let us define two corpora $C(\mathcal{F})$ and $C(\mathcal{E})$ (see figure 1): $C(\mathcal{F})$ is the corpus that consists of a pair (\mathbf{f}, \mathbf{e}) for every sentence $\mathbf{f} \in \mathcal{F}$ and every hypothesis $\mathbf{e} \in L(\mathbf{f})$. Corpus $C(\mathcal{E})$ is defined analogously. Figure 2

E-step_r:

$$\forall \langle \mathbf{f}, \mathbf{e} \rangle \in C(\mathcal{E}): q_r^1(\mathbf{f}, \mathbf{e}) := p_1(\mathbf{f}) \prod_{i=1}^n \frac{\theta_r(f_i, e_i)}{\sum_e \theta_r(f_i, e)}$$

$$\forall \langle \mathbf{f}, \mathbf{e} \rangle \in C(\mathcal{F}): q_r^2(\mathbf{f}, \mathbf{e}) := p_2(\mathbf{e}) \prod_{i=1}^n \frac{\theta_r(f_i, e_i)}{\sum_f \theta_r(f, e_i)}$$

$$\mathbf{M}\text{-step}_r: \text{Define } \mathbb{L}(\mathbf{x}, \mathbf{y}; \theta, p) = p(\mathbf{x}) \prod_{i=1}^n \frac{\theta(x_i, y_i)}{\sum_y \theta(x_i, y)}.$$

$$\theta_{r+1} := \arg \max_{\theta} \sum_{\mathbf{f}, \mathbf{e} \in C(\mathcal{E})} \overbrace{\frac{q_r^1(\mathbf{f}, \mathbf{e})}{Z_r^1(\mathbf{f})} \log \mathbb{L}(\mathbf{f}, \mathbf{e}; \theta, p_1)}^{A_r(\mathbf{f}, \mathbf{e}; \theta)} + \sum_{\mathbf{f}, \mathbf{e} \in C(\mathcal{F})} \overbrace{\frac{q_r^2(\mathbf{f}, \mathbf{e})}{Z_r^2(\mathbf{e})} \log \mathbb{L}(\mathbf{e}, \mathbf{f}; \theta, p_2)}^{B_r(\mathbf{f}, \mathbf{e}; \theta)}$$

Figure 2: Bi-EM algorithm

shows the Bi-EM algorithm, where $Z_r^1(\mathbf{e}) = \sum_{\mathbf{f} \in L(\mathbf{e})} q_r^1(\mathbf{f}, \mathbf{e})$ and $Z_r^2(\mathbf{f}) = \sum_{\mathbf{e} \in L(\mathbf{f})} q_r^2(\mathbf{f}, \mathbf{e})$ are unigram count estimates.

The sum of the two sums in the M-step can be rearranged into a single sum if we precompute a single (complete) corpus \mathcal{C}_r that concatenates $C(\mathcal{F})$ with $C(\mathcal{E})$ and stores the expected frequency counts ($A_r(\mathbf{f}, \mathbf{e}; \theta)$ or $B_r(\mathbf{f}, \mathbf{e}; \theta)$) with each pair as

$$\log \text{freq}_r(\mathbf{f}, \mathbf{e}; \theta) = \begin{cases} A_r(\mathbf{f}, \mathbf{e}; \theta) & \langle \mathbf{f}, \mathbf{e} \rangle \in C(\mathcal{E}) \\ B_r(\mathbf{f}, \mathbf{e}; \theta) & \langle \mathbf{f}, \mathbf{e} \rangle \in C(\mathcal{F}) \end{cases}$$

The M-step becomes the M-step of a standard EM algorithm: $\theta_{r+1} := \arg \max_{\theta} \sum_{\langle \mathbf{f}, \mathbf{e} \rangle \in \mathcal{C}_r} \log \text{freq}_r(\mathbf{f}, \mathbf{e}; \theta)$. Hence, this Bidirectional-EM (Bi-EM) inherits the properties of the common EM algorithm, including convergence and a guarantee of a choice of θ that will not decrease the joint-likelihood after each iteration. The actual update formula is as follows:

$$\overleftarrow{\theta}_r(f, e) = \frac{\sum_{\langle \mathbf{f}, \mathbf{e} \rangle \in \mathcal{C}_r} \log \text{freq}_r(\mathbf{f}, \mathbf{e}; \theta) \times \sum_j \delta[f_j, f] \delta[e_j, e]}{\sum_{\langle \mathbf{f}, \mathbf{e} \rangle \in \mathcal{C}_r} \log \text{freq}_r(\mathbf{f}, \mathbf{e}; \theta)}$$

Note that the Bi-EM takes only twice as much training time as the Uni-EM.

Implementation detail The core of both the Uni-EM estimation methods [7] and the present Bi-EM estimator is the Baum-Welch algorithm [1] for Hidden Markov Models (HMMs), which is known to be an EM algorithm [5]. This algorithm in its most general form employs the Forward-Backward calculations to update expected counts of transition (language model) and emission (lexicon) probabilities. In our setting we fix the language model (transition) estimates and reestimate only the lexicon (emission) probabilities. This is because language models can be readily constructed from large monolingual data and there is no reason to reestimate them.

For the generation of the language models we used the CMU-Cambridge Toolkit [4], employing a first order Markov model. For the Baum-Welch algorithm, we implemented our own (Java) software package. Our software package implements both the Uni- and Bi-EM algorithms².

4 Empirical results

Following [7], our experiments are on noun sequences extracted from corpus sentences. This makes our results comparable to [7]. This experimental setup also separates modeling issues (such as word-order differences, which we do not deal with here) from bi-directional estimation issues (our topic here).

We evaluate different estimators of lexicon probabilities from non-parallel corpora. As an absolute baseline we employ a translation model that assumes uniform lexicon probabilities (called ‘*LM*’ method). The actual baseline, however, is the standard EM [7] (subsequently called *Uni-EM* – Unidirectional EM). We compare these baselines to the present *Bi-EM* algorithm (section 3).

²Bi-EM package is available for download <http://www.still-anonymous-during-submission>

Relatedness	High	Less	Distant/unrelated
German	Europarl-2	European Language News Corpus	Europarl
English	Europarl-1	Gigaword	Gigaword

Table 1: Training corpora and their (relative) relatedness; Europarl-1/2 signifies two different non-parallel portions of Europarl. Three levels of relatedness (high, less, distant).

Training: During training, the input to the estimation methods consists of a non-parallel English-German corpus pair and an ambiguous lexicon³ containing up to seven German translations for every English word.⁴ We initialize the lexicon parameters with a uniform distribution both for Uni- and Bi-EM.

Testing: For evaluation purposes, we embed the lexicon estimates within a simple word-to-word translation system, and evaluate the translation result against the translations available in a given parallel corpus. As [7], we use English and German as the translation language pair. As a test corpus we use 5106 word translation pairs from 1850 noun sequences extracted from an equal number of sentences from the de-news⁵, which have been aligned down to the word level. We measure *accuracy*, the fraction of words whose translation matches the word used in the bitext. In addition, we also provide the *BLEU* scores [12] as an additional measure of translation quality.

4.1 Domain mismatch of source/target corpora

The estimators tested here are expected to operate under a domain- and/or genre-mismatch between the following components (1) source corpus, (2) target corpus, (3) lexicon, and (4) test corpus. Both the lexicon and the test corpus are fixed throughout all experiments. On the one hand, we expect that as the mismatch between the different components becomes more severe, the translation results will degrade. On the other, we would like the *relative* performance differences between the different estimation methods to persist.

Because the joint-likelihood aims at estimates that maximize the joint-likelihood of two corpora, a question may arise as to whether weakening the relatedness (in domain and/or genre) of the two corpora will affect the performance of Bi-EM relative to Uni-EM.

Highly related In a first set of experiments, we use a pair of highly related, *non-parallel* corpora for training purposes. The two corpora consist of noun sequences from two *non-overlapping* sections of the Europarl[6] parallel corpus (English-German). The baseline system using the LM method (uniform lexicon probabilities) achieves an accuracy of 63.11% (BLUE score 0.2372). The accuracy of translation and BLEU score (in parentheses) acquired when using the different estimators, using different training corpora sizes (number of sentence) follows:

#sentences	Uni-EM	Bi-EM
40K	72.01% (0.3896)	76.19% (0.4394)
75K	74.13% (0.4242)	77.34% (0.4660)
100K	74.99% (0.4300)	77.78% (0.4714)

Compared against the baseline (63.11% for the ‘LM’ method) these numbers improve by up to 15% (or in fact 40% error reduction). Bi-EM clearly outperforms the standard Uni-EM. It is evident from the results that the improved accuracy of the Bi-EM does not come from utilizing more data. Bi-EM trained on 40,000 English and the same amount of German sentences significantly outperforms Uni-EM trained on 100,000 English sentences (and a German language model). This is a strong indication that the Joint-Likelihood is a better objective function than the likelihood of a single corpus.

Less related We use as training data newspaper text from the Gigaword (English) and from the European Language Newspaper Text (German), utilizing news stories coming from the same agencies and published during the same period (Associated Press, Agence France-Presse, May 1994-December 1995). Unlike different sections of Europarl, this pair of corpora concerns news texts that originate from non-parallel sources

³The lexicon used by Koehn and Knight is not available (Philipp Koehn, p.c.).

⁴The lexicon was obtained by automatic word alignment of the Europarl corpus.

⁵<http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/de-news/>

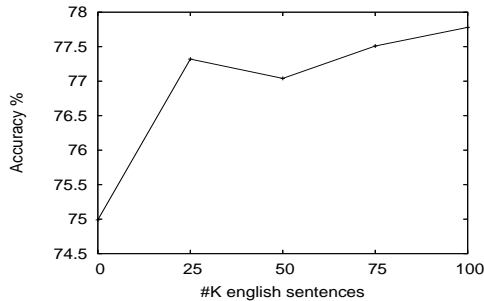


Figure 3: Accuracy of Bi-EM as target training corpus size increases

and are in two different languages. We estimate translation probabilities using Uni-EM and Bi-EM, training with 100K sentences per language used.

#sen	Uni-EM	Bi-EM
100K	70.29% (0.3610)	72.80% (0.3809)

Table 2: Less related: Results

#sen	Uni-EM	Bi-EM
100K	68.90% (0.3110)	70.98% (0.3303)

Table 3: Distantly related: Results

Table 2 shows again that the Bi-EM helps produce significantly more accurate translations. Interestingly, training Bi-EM on 100K sentence still gives better results than Uni-EM trained on 200K sentences (Uni-EM with 200K = 72.08% (0.3737)).

Distantly related We also trained on a pair of distantly related corpora. These are the newspaper text from Gigaword (English) and the parliament proceedings from Europarl (German). As seen in table 3, Bi-EM is still able to produce estimates that give more accurate translations than Uni-EM. Again Bi-EM trained on 100K sentences outperforms Uni-EM trained on 200K sentences (Uni-EM on 200K = 70.23% (0.3215)).

Smaller target language data The experiments investigate (1) the effect of having source and target corpora of different sizes, which is common, especially when one side is from a resource-poor language, and (2) the utility of joint-likelihood estimation (maximize likelihood of source and target corpora), as opposed to standard likelihood maximization of the source corpus alone.

We employ the same corpora as in section 4.1, varying this time the amount of training sentences from the target language (English), while maintaining a training corpus of 100K German sentences in all cases. Figure 3 shows the average accuracies of Bi-EM as function of target corpus increase. Note that the zero point refers to the Bi-EM trained on target corpus of size zero, which is equivalent to the Uni-EM. Interestingly, 81% of the accuracy increase of Bi-EM relative to Uni-EM is already obtained by using only 25K sentences, 77.32% (0.4542). These accuracies are averages over 3 different non-overlapping sets of 25K English sentences. This result shows that Bi-EM is more powerful than Uni-EM even if one of the two corpora is small. This confirms our hypothesis that joint-likelihood estimation (i.e., Bi-EM) is more effective than maximizing the likelihood of one side of the translation channel.

5 Conclusions

Maximizing the joint-likelihood of a pair of source and target corpora as an objective function gives better translation lexicon estimates than maximizing the likelihood of the source corpus alone. While the joint-likelihood estimator is based on a non-directional, joint probability model embedded in two directional translation models, the standard approach is based on a single, conditional, directional translation model. We have shown in this paper how the joint-likelihood can be optimized using a bi-directional EM algorithm.

Our extensive experiments show the utility of maximizing the joint likelihood of the source and target corpora when training a translation lexicon. The Bi-EM gives better estimates with much less data. Crucially, the Bi-EM delivers better results than the Uni-EM regardless of mismatch in domain or genre between

the source and target corpora. The estimates of all algorithms give less accurate translation as the source and target training corpora become less related. This is expected because the less related the source and target corpora, the less of a translation relation exists between the words of the two corpora.

In future work we aim at utilizing the Bi-EM for porting linguistic processing tools from a resource-rich to a resource-poor language in cases where there exist no parallel corpora. We think that the Bi-EM could be useful in statistical machine translation, in particular for obtaining better alignments, as [14] have shown, but also in obtaining improved translation model estimates. Whenever a joint channel model is postulated and data from source and target sides is available, it makes more sense to employ Bi-EM than standard Uni-directional EM.

References

- [1] L.E. Baum, T. Peterie, G. Souled, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41(1):164–171, 1970.
- [2] P. Brown, J. Cocke, S. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A statistical approach to language translation. In *COLING-88*, 1988.
- [3] David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. Parsing arabic dialects. In *EACL*. The Association for Computer Linguistics, 2006.
- [4] P.R. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings ESCA Eurospeech*, 1997.
- [5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
- [6] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*, 2005.
- [7] P. Koehn and K. Knight. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *AAAI/IAAI*, 2000.
- [8] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003)*, Edmonton, Canada, May 2003.
- [9] P. Liang, B. Taskar, and D. Klein. Alignment by agreement. In *Proceedings of the Human Language Technology Conference (HLT-NAACL 2006)*, New York, June 2006.
- [10] M. Mylonakis, K. Sima'an, and R. Hwa. Unsupervised estimation for noisy-channel models. In *Proceedings of International Conference on Machine Learning (ICML'07)*, Michigan, USA, 2007.
- [11] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March 2003.
- [12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, Morristown, NJ, USA, 2001. Association for Computational Linguistics.
- [13] O. Rambow, D. Chiang, M. Diab, N. Habash, R. Hwa, K. Sima'an, V. Lacey, R. Levy, C. Nichols, and S. Shareef. Parsing arabic dialects. Technical report, Johns Hopkins University 2005 Summer Workshop on Language Engineering, 2005.
- [14] R. Zens, E. Matusov, and H. Ney. Improved word alignment using a symmetric lexicon model. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing)*, pages 36–42, Geneva, Switzerland, August 2004.