Toward Statistical Machine Translation without Parallel Corpora

Alexandre Klementiev **Ann Irvine** **Chris Callison-Burch** **David Yarowsky**
Center for Language and Speech Processing
Johns Hopkins University

Abstract

We estimate the parameters of a phrase-based statistical machine translation system from *monolingual* corpora instead of a *bilingual* parallel corpus. We extend existing research on bilingual lexicon induction to estimate *both* lexical and phrasal translation probabilities for MT-scale phrase-tables. We propose a novel algorithm to estimate reordering probabilities from monolingual data. We report translation results for an end-to-end translation system using these monolingual features alone. Our method only requires monolingual corpora in source and target languages, a small bilingual dictionary, and a small bitext for tuning feature weights. In this paper, we examine an idealization where a phrase-table is given. We examine the degradation in translation performance when bilingually estimated translation probabilities are removed and show that 80%+ of the loss can be recovered with monolingually estimated features alone. We further show that our monolingual features add 1.5 BLEU points when combined with standard bilingually estimated phrase table features.

1 Introduction

The parameters of statistical models of translation are typically estimated from large bilingual parallel corpora (Brown et al., 1993). However, these resources are not available for most language pairs, and they are expensive to produce in quantities sufficient for building a good translation system (Germann, 2001). We attempt an entirely different approach; we use cheap and plentiful monolingual resources to induce an end-to-end statistical machine translation system. In particular, we extend the long line of work on inducing translation lexicons (beginning with Rapp (1995)) and propose to use multiple independent cues present in monolingual texts to estimate lexical and phrasal translation probabilities for large, MT-scale phrase-tables. We then introduce a novel algorithm to estimate reordering features from monolingual data alone, and we report the performance of a phrase-based statistical model (Koehn et al., 2003) estimated using these monolingual features.

Most of the prior work on lexicon induction is motivated by the idea that it could be applied to machine translation but stops short of actually doing so. Lexicon induction holds the potential to create machine translation systems for languages which do not have extensive parallel corpora. Training would only require two large monolingual corpora and a small bilingual dictionary, if one is available. The idea is that intrinsic properties of monolingual data (possibly along with a handful of bilingual pairs to act as example mappings) can provide independent but informative cues to learn translations because words (and phrases) behave similarly across languages. This work is the first attempt to extend and apply these ideas to an end-to-end machine translation pipeline. While we make an explicit assumption that a table of phrasal translations is given a priori, we induce every other parameter of a full phrase-based translation system from monolingual data alone. The contributions of this work are:

- In Section 2.2 we analyze the challenges of using bilingual lexicon induction for statistical MT (performance on low frequency items, and moving from words to phrases).
- In Sections 3.1 and 3.2 we use multiple cues present in monolingual data to estimate lexical and phrasal translation scores.
- In Section 3.3 we propose a novel algorithm for estimating phrase reordering features from monolingual texts.
- Finally, in Section 5 we systematically drop feature functions from a phrase table and then replace them with monolingually estimated equivalents, reporting end-to-end translation quality.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 130–140,
Avignon, France, April 23 - 27 2012.

2 Background

We begin with a brief overview of the standard phrase-based statistical machine translation model. Here, we define the parameters which we later replace with monolingual alternatives. We continue with a discussion of bilingual lexicon induction; we extend these methods to estimate the monolingual parameters in Section 3. This approach allows us to replace expensive/rare bilingual parallel training data with two large monolingual corpora, a small bilingual dictionary, and ≈2,000 sentence bilingual development set, which are comparatively plentiful/inexpensive.

2.1 Parameters of phrase-based SMT

Statistical machine translation (SMT) was first formulated as a series of probabilistic models that learn word-to-word correspondences from sentence-aligned bilingual parallel corpora (Brown et al., 1993). Current methods, including *phrase-based* (Och, 2002; Koehn et al., 2003) and *hierarchical* models (Chiang, 2005), typically start by word-aligning a bilingual parallel corpus (Och and Ney, 2003). They extract multi-word phrases that are consistent with the Viterbi word alignments and use these phrases to build new translations. A variety of parameters are estimated using the bitexts. Here we review the parameters of the standard phrase-based translation model (Koehn et al., 2007). Later we will show how to estimate them using monolingual texts instead. These parameters are:

- *Phrase pairs*. Phrase extraction heuristics (Venugopal et al., 2003; Tillmann, 2003; Och and Ney, 2004) produce a set of phrase pairs (e, f) that are consistent with the word alignments. In this paper we assume that the phrase pairs are given (without any scores), and we induce every other parameter of the phrase-based model from monolingual data.

- *Phrase translation probabilities*. Each phrase pair has a list of associated feature functions (FFs). These include phrase translation probabilities, $\phi(e|f)$ and $\phi(f|e)$, which are typically calculated via maximum likelihood estimation.

- *Lexical weighting*. Since MLE overestimates ϕ for phrase pairs with sparse counts, lexical weighting FFs are used to smooth. Average word translation probabilities, $w(e_i|f_j)$, are calculated via phrase-pair-internal word alignments.

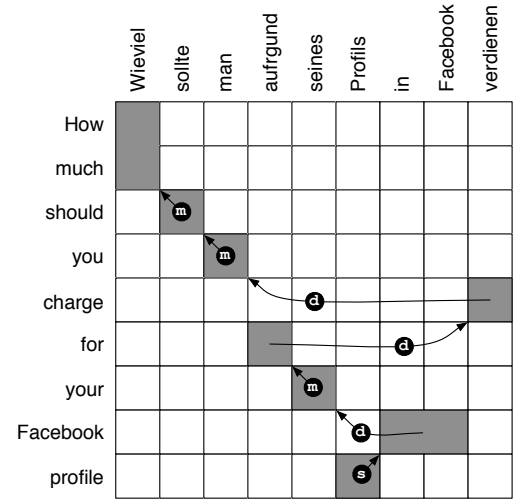

Figure 1: The reordering probabilities from the phrase-based models are estimated from bilingual data by calculating how often in the parallel corpus a phrase pair (f, e) is orientated with the preceding phrase pair in the 3 types of orientations (monotone, swapped, and discontinuous).

- *Reordering model*. Each phrase pair (e, f) also has associated reordering parameters, $p_o(\text{orientation}|f, e)$, which indicate the distribution of its orientation with respect to the previously translated phrase. Orientations are *monotone, swap, discontinuous* (Tillman, 2004; Kumar and Byrne, 2004), see Figure 1.

- *Other features*. Other typical features are n-gram language model scores and a phrase penalty, which governs whether to use fewer longer phrases or more shorter phrases. These are not bilingually estimated, so we can re-use them directly without modification.

The features are combined in a log linear model, and their weights are set through minimum error rate training (Och, 2003). We use the same log linear formulation and MERT but propose alternatives derived directly from monolingual data for all parameters except for the phrase pairs themselves. Our pipeline still requires a small bitext of approximately 2,000 sentences to use as a development set for MERT parameter tuning.

2.2 Bilingual lexicon induction for SMT

Bilingual lexicon induction describes the class of algorithms that attempt to learn translations from monolingual corpora. Rapp (1995) was the first to propose using non-parallel texts to learn the translations of words. Using large, unrelated English and German corpora (with 163m and 135m words) and a small German-English bilingual dictionary (with 22k entires), Rapp (1999) demonstrated that reasonably accurate translations could be learned for 100 German nouns that were not contained in the seed bilingual dictionary. His algorithm worked by (1) building a context vector representing an unknown German word by counting its co-occurrence with all the other words in the German monolingual corpus, (2) projecting this German vector onto the vector space of English using the seed bilingual dictionary, (3) calculating the similarity of this sparse projected vector to vectors for English words that were constructed using the English monolingual corpus, and (4) outputting the English words with the highest similarity as the most likely translations.

A variety of subsequent work has extended the original idea either by exploring different measures of vector similarity (Fung and Yee, 1998) or by proposing other ways of measuring similarity beyond co-occurence within a context window. For instance, Schafer and Yarowsky (2002) demonstrated that word translations tend to co-occur *in time* across languages. Koehn and Knight (2002) used similarity *in spelling* as another kind of cue that a pair of words may be translations of one another. Garera et al. (2009) defined context vectors using *dependency relations* rather than adjacent words. Bergsma and Van Durme (2011) used the *visual similarity* of labeled web images to learn translations of nouns. Additional related work on learning translations from monolingual corpora is discussed in Section 6.

In this paper, we apply bilingual lexicon induction methods to statistical machine translation. Given the obvious benefits of not having to rely on scarce bilingual parallel training data, it is surprising that bilingual lexicon induction has not been used for SMT before now. There are several open questions that make its applicability to SMT uncertain. Previous research on bilingual lexicon induction learned translations only for a small number of high frequency words (e.g. 100 nouns in Rapp (1995), 1,000 most frequent words in Koehn and Knight (2002), or 2,000 most frequent nouns in Haghighi et al. (2008)). Although previous work reported high translation accuracy, it may be misleading to extrapolate the results to SMT, where it is necessary to translate a much larger set of words and phrases, including many low frequency items.

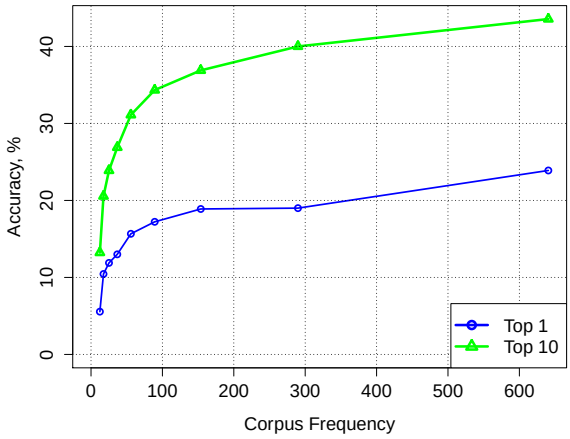

Figure 2: Accuracy of single-word translations induced using contextual similarity as a function of the source word corpus frequency. Accuracy is the proportion of the source words with at least one correct (bilingual dictionary) translation in the top 1 and top 10 candidate lists.

In a preliminary study, we plotted the accuracy of translations against the frequency of the source words in the monolingual corpus. Figure 2 shows the result for translations induced using contextual similarity (defined in Section 3.1). Unsurprisingly, frequent terms have a substantially better chance of being paired with a correct translation, with words that only occur once having a low chance of being translated accurately.[1] This problem is exacerbated when we move to multi-token phrases. As with phrase translation features estimated from parallel data, longer phrases are more sparse, making similarity scores less reliable than for single words.

Another impediment (not addressed in this paper) for using lexicon induction for SMT is the number of translations that must be learned. Learning translations for all words in the source language requires n^2 vector comparisons, since each word in the source language vocabulary must

[1] For a description of the experimental setup used to produce these translations, see Experiment 8 in Section 5.2.

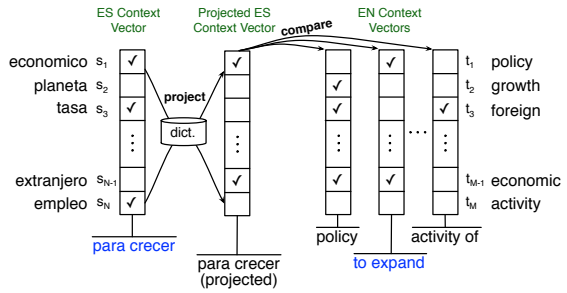

Figure 3: Scoring contextual similarity of *phrases*: first, contextual vectors are projected using a small seed dictionary and then compared with the target language candidates.

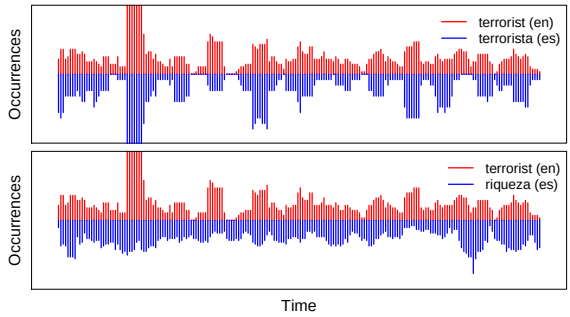

Figure 4: Temporal histograms of the English phrase *terrorist*, its Spanish translation *terrorista*, and *riqueza* (wealth) collected from monolingual texts spanning a 13 year period. While the correct translation has a good temporal match, the non-translation *riqueza* has a distinctly different signature.

be compared against the vectors for all words in the target language vocabulary. The size of the n^2 comparisons hugely increases if we compare vectors for multi-word phrases instead of just words. In this work, we avoid this problem by assuming that a limited set of phrase pairs is given a priori (but without scores). By limiting ourselves to phrases in a phrase table, we vastly limit the search space of possible translations. This is an idealization because high quality translations are guaranteed to be present. However, as our lesion experiments in Section 5.1 show, a phrase table without accurate translation probability estimates is insufficient to produce high quality translations. We show that lexicon induction methods can be used to replace bilingual estimation of phrase- and lexical-translation probabilities, making a significant step towards SMT without parallel corpora.

3 Monolingual Parameter Estimation

We use bilingual lexicon induction methods to estimate the parameters of a phrase-based translation model from monolingual data. Instead of scores estimated from bilingual parallel data, we make use of cues present in monolingual data to provide multiple orthogonal estimates of similarity between a pair of phrases.

3.1 Phrasal similarity features

Contextual similarity. We extend the vector space approach of Rapp (1999) to compute similarity between *phrases* in the source and target languages. More formally, assume that $(s_1, s_2, \ldots s_N)$ and $(t_1, t_2, \ldots t_M)$ are (arbitrarily indexed) source and target vocabularies, respectively. A source phrase f is represented with an N- and target phrase e with an M-dimensional vector (see Figure 3). The component values of the vector representing a phrase correspond to how often each of the words in that vocabulary appear within a two word window on either side of the phrase. These counts are collected using monolingual corpora. After the values have been computed, a contextual vector f is projected onto the English vector space using translations in a seed bilingual dictionary to map the component values into their appropriate English vector positions. This sparse projected vector is compared to the vectors representing all English phrases e. Each phrase pair in the phrase table is assigned a contextual similarity score $c(f, e)$ based on the similarity between e and the projection of f.

Various means of computing the component values and vector similarity measures have been proposed in literature (e.g. Rapp (1999), Fung and Yee (1998)). Following Fung and Yee (1998), we compute the value of the k-th component of f's contextual vector as follows:

$$w_k = n_{f,k} \times (log(n/n_k) + 1)$$

where $n_{f,k}$ and n_k are the number of times s_k appears in the context of f and in the entire corpus, and n is the maximum number of occurrences of any word in the data. Intuitively, the more frequently s_k appears with f and the less common it is in the corpus in general, the higher its component value. Similarity between two vectors is measured as the cosine of the angle between them.

Temporal similarity. In addition to contextual similarity, phrases in two languages may

be scored in terms of their temporal similarity (Schafer and Yarowsky, 2002; Klementiev and Roth, 2006; Alfonseca et al., 2009). The intuition is that news stories in different languages will tend to discuss the same world events on the same day. The frequencies of translated phrases over time give them particular signatures that will tend to spike on the same dates. For instance, if the phrase *asian tsunami* is used frequently during a particular time span, the Spanish translation *maremoto asiático* is likely to also be used frequently during that time. Figure 4 illustrates how the temporal distribution of *terrorist* is more similar to Spanish *terrorista* than to other Spanish phrases. We calculate the temporal similarity between a pair of phrases $t(f, e)$ using the method defined by Klementiev and Roth (2006). We generate a temporal signature for each phrase by sorting the set of (time-stamped) documents in the monolingual corpus into a sequence of equally sized temporal bins and then counting the number of phrase occurrences in each bin. In our experiments, we set the window size to 1 day, so the size of temporal signatures is equal to the number of days spanned by our corpus. We use cosine distance to compare the normalized temporal signatures for a pair of phrases (f, e).

Topic similarity. Phrases and their translations are likely to appear in articles written about the same topic in two languages. Thus, topic or category information associated with monolingual data can also be used to indicate similarity between a phrase and its candidate translation. In order to score a pair of phrases, we collect their topic signatures by counting their occurrences in each topic and then comparing the resulting vectors. We again use the cosine similarity measure on the normalized topic signatures. In our experiments, we use interlingual links between Wikipedia articles to estimate topic similarity. We treat each linked article pair as a topic and collect counts for each phrase across all articles in its corresponding language. Thus, the size of a phrase topic signature is the number of article pairs with interlingual links in Wikipedia, and each component contains the number of times the phrase appears in (the appropriate side of) the corresponding pair. Our Wikipedia-based topic similarity feature, $w(f, e)$, is similar in spirit to polylingual topic models (Mimno et al., 2009), but it is scalable to full bilingual lexicon induction.

3.2 Lexical similarity features

In addition to the three phrase similarity features used in our model – $c(f, e), t(f, e)$ and $w(f, e)$ – we include four additional *lexical similarity features* for each of phrase pair. The first three lexical features $c_{lex}(f, e), t_{lex}(f, e)$ and $w_{lex}(f, e)$ are the lexical equivalents of the phrase-level *contextual, temporal* and *wikipedia topic* similarity scores. They score the similarity of individual words within the phrases. To compute these lexical similarity features, we average similarity scores over all possible word alignments across the two phrases. Because individual words are more frequent than multiword phrases, the accuracy of c_{lex}, t_{lex}, and w_{lex} tends to be higher than their phrasal equivalents (this is similar to the effect observed in Figure 2).

Orthographic / phonetic similarity. The final lexical similarity feature that we incorporate is $o(f, e)$, which measures the orthographic similarity between words in a phrase pair. Etymologically related words often retain similar spelling across languages with the same writing system, and low string edit distance sometimes signals translation equivalency. Berg-Kirkpatrick and Klein (2011) present methods for learning correspondences between the alphabets of two languages. We can also extend this idea to language pairs not sharing the same writing system since many cognates, borrowed words, and names remain phonetically similar. Transliterations can be generated for tokens in a source phrase (Knight and Graehl, 1997), with $o(f, e)$ calculating phonetic similarity rather than orthographic.

The three phrasal and four lexical similarity scores are incorporated into the log linear translation model as feature functions, replacing the bilingually estimated phrase translation probabilities ϕ and lexical weighting probabilities w. Our seven similarity scores are not the only ones that could be incorporated into the translation model. Various other similarity scores can be computed depending on the available monolingual data and its associated metadata (see, e.g. Schafer and Yarowsky (2002)).

3.3 Reordering

The remaining component of the phrase-based SMT model is the reordering model. We introduce a novel algorithm for estimating

Input: Source and target phrases f and e,
Source and target monolingual corpora C_f and C_e,
Phrase table pairs $T = \{(f^{(i)}, e^{(i)})\}_{i=1}^{N}$.
Output: Orientation features (p_m, p_s, p_d).

$S_f \leftarrow$ sentences containing f in C_f;
$S_e \leftarrow$ sentences containing e in C_e;
$(B_f, -, -) \leftarrow$ CollectOccurs$(f, \cup_{i=1}^{N} f^{(i)}, S_f)$;
$(B_e, A_e, D_e) \leftarrow$ CollectOccurs$(e, \cup_{i=1}^{N} e^{(i)}, S_e)$;
$c_m = c_s = c_d = 0$;
foreach *unique* f' *in* B_f **do**
  foreach *translation* e' *of* f' *in* T **do**
    $c_m = c_m + \#_{B_e}(e')$;
    $c_s = c_s + \#_{A_e}(e')$;
    $c_d = c_d + \#_{D_e}(e')$;
$c \leftarrow c_m + c_s + c_d$;
return $(\frac{c_m}{c}, \frac{c_s}{c}, \frac{c_d}{c})$

CollectOccurs(*r, R, S*)
  $B \leftarrow (); A \leftarrow (); D \leftarrow ();$
  foreach *sentence* $s \in S$ **do**
    foreach *occurrence of phrase* r *in* s **do**
      $B \leftarrow B$ + (longest preceding r and in R);
      $A \leftarrow A$ + (longest following r and in R);
      $D \leftarrow D$ + (longest discontinuous w/ r and in R);
  return *(B, A, D)*;

Figure 5: Algorithm for estimating reordering probabilities from monolingual data.

$p_o(\text{orientation}|f, e)$ from two monolingual corpora instead a bitext.

Figure 1 illustrates how the phrase pair orientation statistics are estimated in the standard phrase-based SMT pipeline. For a phrase pair like (f = *"Profils"*, e = *"profile"*), we count its orientation with the previously translated phrase pair (f' = *"in Facebook"*, e' = *"Facebook"*) across all translated sentence pairs in the bitext.

In our pipeline we do not have translated sentence pairs. Instead, we look for monolingual sentences in the source corpus which contain the source phrase that we are interested in, like f = *"Profils"*, and at least one other phrase that we have a translation for, like f' = *"in Facebook"*. We then look for all target language sentences in the target monolingual corpus that contain the translation of f (here e = *"profile"*) and any translation of f'. Figure 6 illustrates that it is possible to find evidence for $p_o(\text{swapped}|Profils, profile)$, even from the non-parallel, non-translated sentences drawn from two independent monolingual corpora. By looking for foreign sentences containing pairs of adjacent foreign phrases (f, f') and English sentences con-

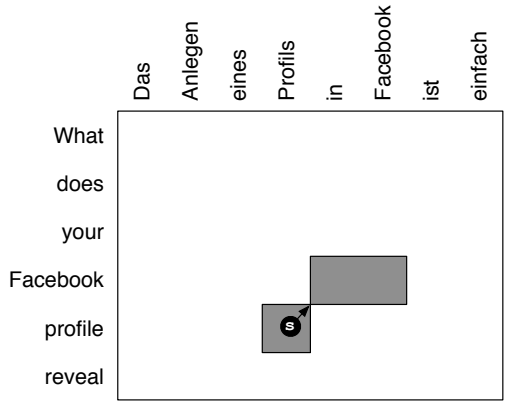

Figure 6: Collecting phrase orientation statistics for a English-German phrase pair (*"profile"*, *"Profils"*) from non-parallel sentences (the German sentence translates as *"Creating a Facebook profile is easy"*).

taining their corresponding translations (e, e'), we are able to increment orientation counts for (f, e) by looking at whether e and e' are adjacent, swapped, or discontinuous. The orientations correspond directly to those shown in Figure 1.

One subtly of our method is that shorter and more frequent phrases (e.g. punctuation) are more likely to appear in multiple orientations with a given phrase, and therefore provide poor evidence of reordering. Therefore, we (a) collect the longest contextual phrases (which also appear in the phrase table) for reordering feature estimation, and (b) prune the set of sentences so that we only keep a small set of least frequent contextual phrases (this has the effect of dropping many function words and punctuation marks and and relying more heavily on multi-word content phrases to estimate the reordering).[2]

Our algorithm for learning the reordering parameters is given in Figure 5. The algorithm estimates a probability distribution over monotone, swap, and discontinuous orientations (p_m, p_s, p_d) for a phrase pair (f, e) from two monolingual corpora C_f and C_e. It begins by calling `CollectOccurs` to collect the longest matching phrase table phrases that precede f in source monolingual data (B_f), as well as those that precede (B_e), follow (A_e), and are discontinuous (D_e) with e in the target language data. For each unique phrase f' preceding f, we look up translations in the phrase table *T*. Next, we count[3] how

[2]The pruning step has an additional benefit of minimizing the memory needed for orientation feature estimations.

[3]$\#_L(x)$ returns the count of object x in list L.

Monolingual training corpora			
	Europarl	Gigaword	Wikipedia
date range	4/96-10/09	5/94-12/08	n/a
uniq shared dates	829	5,249	n/a
Spanish articles	n/a	3,727,954	59,463
English articles	n/a	4,862,876	59,463
Spanish lines	1,307,339	22,862,835	2,598,269
English lines	1,307,339	67,341,030	3,630,041
Spanish words	28,248,930	774,813,847	39,738,084
English words	27,335,006	1,827,065,374	61,656,646

Spanish-English phrase table	
Phrase pairs	3,093,228
Spanish phrases	89,386
English phrases	926,138
Spanish unigrams	13,216
Avg # translations	98.7
Spanish bigrams	41,426
Avg # translations	31.9
Spanish trigrams	34,744
Avg # translations	13.5

Table 1: Statistics about the monolingual training data and the phrase table that was used in all of the experiments.

many translations e' of f' appeared before, after or were discontinuous with e in the target language data. Finally, the counts are normalized and returned. These normalized counts are the values we use as estimates of $p_o(\text{orientation}|f, e)$.

4 Experimental Setup

We use the Spanish-English language pair to test our method for estimating the parameters of an SMT system from monolingual corpora. This allows us to compare our method against the normal bilingual training procedure. We expect bilingual training to result in higher translation quality because it is a more direct method for learning translation probabilities. We systematically remove different parameters from the standard phrase-based model, and then replace them with our monolingual equivalents. Our goal is to recover as much of the loss as possible for each of the deleted bilingual components.

The standard phrase-based model that we use as our top-line is the Moses system (Koehn et al., 2007) trained over the full Europarl v5 parallel corpus (Koehn, 2005). With the exception of maximum phrase length (set to 3 in our experiments), we used default values for all of the parameters. All experiments use a trigram language model trained on the English side of the Europarl corpus using SRILM with Kneser-Ney smoothing. To tune feature weights in minimum error rate training, we use a development bitext of 2,553 sentence pairs, and we evaluate performance on a test set of 2,525 single-reference translated newswire articles. These development and test datasets were distributed in the WMT shared task (Callison-Burch et al., 2010).[4] MERT was re-run for every experiment.

We estimate the parameters of our model from two sets of monolingual data, detailed in Table 1:

- First, we treat the two sides of the Europarl parallel corpus as independent, monolingual corpora. Haghighi et al. (2008) also used this method to show how well translations could be learned from monolingual corpora under ideal conditions, where the contextual and temporal distribution of words in the two monolingual corpora are nearly identical.

- Next, we estimate the features from truly monolingual corpora. To estimate the *contextual* and *temporal* similarity features, we use the Spanish and English Gigaword corpora.[5] These corpora are substantially larger than the Europarl corpora, providing 27x as much Spanish and 67x as much English for contextual similarity, and 6x as many paired dates for temporal similarity. *Topical* similarity is estimated using Spanish and English Wikipedia articles that are paired with interlanguage links.

To project context vectors from Spanish to English, we use a bilingual dictionary containing entries for 49,795 Spanish words. Note that end-to-end translation quality is robust to substantially reducing dictionary size, but we omit these experiments due to space constraints. The context vectors for words and phrases incorporate co-occurrence counts using a two-word window on either side.

The title of our paper uses the word *towards* because we assume that an inventory of phrase pairs is given. Future work will explore inducing the

[4] Specifcially, *news-test2008* plus *news-syscomb2009* for dev and *newstest2009* for test.

[5] We use the afp, apw and xin sections of the corpora.

	Exp	Phrase scores / orientation scores
1	B/B	bilingual / bilingual (Moses)
2	B/-	bilingual / distortion
3	-/B	none / bilingual
4	-/-	none / distortion
5, 12	-/M	none / mono
6, 13	t/-	temporal mono / distortion
7,14	o/-	orthographic mono / distortion
8, 15	c/-	contextual mono / distortion
16	w/-	Wikipedia topical mono / distorion
9, 17	M/-	all mono / distortion
10, 18	M/M	all mono / mono
11, 19	BM/B	bilingual + all mono / bilingual

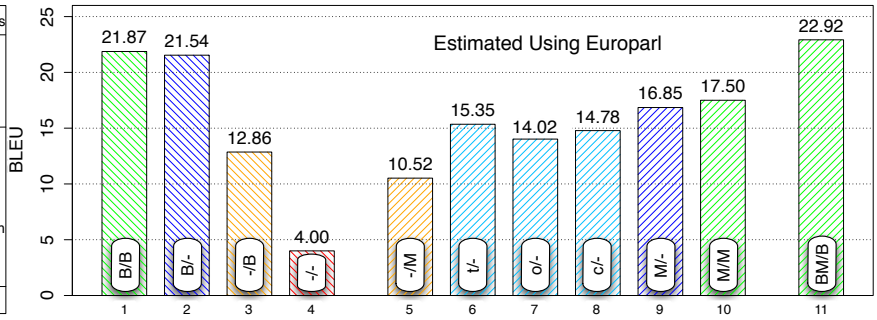

Figure 7: Much of the loss in BLEU score when bilingually estimated features are removed from a Spanish-English translation system (experiments 1-4) can be recovered when they are replaced with monolingual equivalents estimated from monolingual Europarl data (experiments 5-10). The labels indicate how the different types of parameters are estimated, the first part is for phrase-table features, the second is for reordering probabilities.

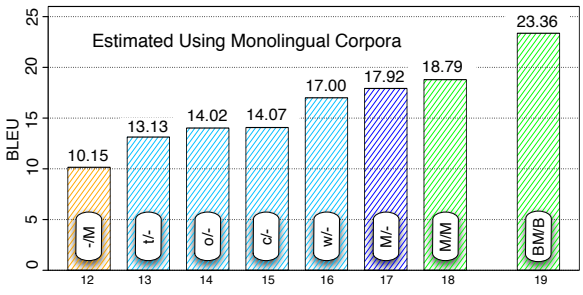

Figure 8: Performance of monolingual features derived from truly monolingual corpora. Over 82% of the BLEU score loss can be recovered.

phrase table itself from monolingual texts. Across all of our experiments, we use the phrase table that the bilingual model learned from the Europarl parallel corpus. We keep its phrase pairs, but we drop all of its scores. Table 1 gives details of the phrase pairs. In our experiments, we estimated similarity and reordering scores for more than 3 million phrase pairs. For each source phrase, the set of possible translations was constrained and likely to contain good translations. However, the average number of possible translations was high (ranging from nearly 100 translations for each unigram to 14 for each trigram). These contain a lot of noise and result in low end-to-end translation quality without good estimates of translation quality, as the experiments in Section 5.1 show.

Software. Because many details of our estimation procedures must be omitted for space, we distribute our full set of code along with scripts for running our experiments and output translations. These may be downed from `http://www.cs.jhu.edu/~anni/papers/lowresmt/`

5 Experimental Results

Figures 7 and 8 give experimental results. Figure 7 shows the performance of the standard phrase-based model when each of the bilingually estimated features are removed. It shows how much of the performance loss can be recovered using our monolingual features when they are estimated from the Europarl training corpus but treating each side as an independent, monolingual corpus. Figure 8 shows the recovery when using truly monolingual corpora to estimate the parameters.

5.1 Lesion experiments

Experiments 1-4 remove bilingually estimated parameters from the standard model. For Spanish-English, the relative contribution of the phrase-table features (which include the phrase translation probabilities ϕ and the lexical weights w) is greater than the reordering probabilities. When the reordering probability $p_o(\text{orientation}|f,e)$ is eliminated and replaced with a simple distance-based distortion feature that does not require a bitext to estimate, the score dips only marginally since word order in English and Spanish is similar. However, when both the reordering and the phrase table features are dropped, leaving only the LM feature and the phrase penalty, the resulting translation quality is abysmal, with the score dropping a total of over 17 BLEU points.

5.2 Adding equivalent monolingual features estimated using Europarl

Experiments 5-10 show how much our monolingual equivalents could recover when the monolingual corpora are drawn from the two sides of the bitext. For instance, our algorithm for estimating

reordering probabilities from monolingual data (**–/M**) adds 5 BLEU points, which is 73% of the potential recovery going from the model (**–/–**) to the model with bilingual reordering features (**–/B**).

Of the **t**emporal, **o**rthographic, and **c**ontextual monolingual features the temporal feature performs the best. Together (**M/–**), they recover more than each individually. Combining monolingually estimated reordering and phrase table features (**M/M**) yields a total gain of 13.5 BLEU points, or over 75% of the BLEU score loss that occurred when we dropped all features from the phrase table. However, these results use "monolingual" corpora which have practically identical phrasal and temporal distributions.

5.3 Estimating features using *truly monolingual* corpora

Experiments 12-18 estimate all of the features from truly monolingual corpora. Our novel algorithm for estimating reordering holds up well and recovers 69% of the loss, only 0.4 BLEU points less than when estimated from the Europarl monolingual texts. The temporal similarity feature does not perform as well as when it was estimated using Europarl data, but the contextual feature does. The topic similarity using Wikipedia performs the strongest of the individual features.

Combining the monolingually estimated reordering features with the monolingually estimated similarity features (**M/M**) yields a total gain of **14.8 BLEU** points, or **over 82%** of the BLEU point loss that occurred when we dropped all features from the phrase table. This is equivalent to training the standard system on a bitext with roughly 60,000 lines or nearly 2 million words (learning curve omitted for space).

Finally, we supplement the standard bilingually estimated model parameters with our monolingual features (**BM/B**), and we see a **1.5 BLEU** point increase over the standard model. Therefore, our monolingually estimated scores capture some novel information not contained in the standard feature set.

6 Additional Related Work

Carbonell et al. (2006) described a data-driven MT system that used no parallel text. It produced translation lattices using a bilingual dictionary and scored them using an n-gram language model. Their method has no notion of translation similarity aside from a bilingual dictionary. Similarly, Sánchez-Cartagena et al. (2011) supplement an SMT phrase table with translation pairs extracted from a bilingual dictionary and give each a frequency of one for computing translation scores.

Ravi and Knight (2011) treat MT without parallel training data as a decipherment task and learn a translation model from monolingual text. They translate corpora of Spanish time expressions and subtitles, which both have a limited vocabulary, into English. Their method has not been applied to broader domains of text.

Most work on learning translations from monolingual texts only examine small numbers of frequent words. Huang et al. (2005) and Daumé and Jagarlamudi (2011) are exceptions that improve MT by mining translations for OOV items.

A variety of past research has focused on mining parallel or comparable corpora from the web (Munteanu and Marcu, 2006; Smith et al., 2010; Uszkoreit et al., 2010). Others use an existing SMT system to discover parallel sentences within independent monolingual texts, and use them to re-train and enhance the system (Schwenk, 2008; Chen et al., 2008; Schwenk and Senellart, 2009; Rauf and Schwenk, 2009; Lambert et al., 2011). These are complementary but orthogonal to our research goals.

7 Conclusion

This paper has demonstrated a novel set of techniques for successfully estimating phrase-based SMT parameters from *monolingual* corpora, potentially circumventing the need for large bitexts, which are expensive to obtain for new languages and domains. We evaluated the performance of our algorithms in a full end-to-end translation system. Assuming that a bilingual-corpus-derived phrase table is available, we were able utilize our monolingually-estimated features to recover over 82% of BLEU loss that resulted from removing the bilingual-corpus-derived phrase-table probabilities. We also showed that our monolingual features add 1.5 BLEU points when combined with standard bilingually estimated features. Thus our techniques have stand-alone efficacy when large bilingual corpora are not available and also make a significant contribution to combined ensemble performance when they are.

References

Enrique Alfonseca, Massimiliano Ciaramita, and Keith Hall. 2009. Gazpacho and summer rash: lexical relationships from temporal patterns of web search queries. In *Proceedings of EMNLP*.

Taylor Berg-Kirkpatrick and Dan Klein. 2011. Simple effective decipherment via combinatorial optimization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP-2011)*, Edinburgh, Scotland, UK.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Peter Brown, John Cocke, Stephen Della Pietra, Vincent Della Pietra, Frederick Jelinek, Robert Mercer, and Paul Poossin. 1988. A statistical approach to language translation. In *12th International Conference on Computational Linguistics (CoLing-1988)*.

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*.

Jaime Carbonell, Steve Klein, David Miller, Michael Steinbaum, Tomer Grassiany, and Jochen Frey. 2006. Context-based machine translation. In *Proceedings of AMTA*.

Boxing Chen, Min Zhang, Aiti Aw, and Haizhou Li. 2008. Exploiting n-best hypotheses for SMT self-enhancement. In *Proceedings of ACL/HLT*, pages 157–160.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*.

Hal Daumé and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of ACL/HLT*.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of ACL/CoLing*.

Nikesh Garera, Chris Callison-Burch, and David Yarowsky. 2009. Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Thirteenth Conference On Computational Natural Language Learning (CoNLL-2009)*, Boulder, Colorado.

Ulrich Germann. 2001. Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In *ACL 2001 Workshop on Data-Driven Machine Translation*, Toulouse, France.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *Proceedings of ACL/HLT*.

Fei Huang, Ying Zhang, and Stephan Vogel. 2005. Mining key phrase translations from web corpora. In *Proceedings of EMNLP*.

Alexandre Klementiev and Dan Roth. 2006. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the ACL/Coling*.

Kevin Knight and Jonathan Graehl. 1997. Machine transliteration. In *Proceedings of ACL*.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *ACL Workshop on Unsupervised Lexical Acquisition*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL-2007 Demo and Poster Sessions*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*.

Shankar Kumar and William Byrne. 2004. Local phrase reordering models for statistical machine translation. In *Proceedings of HLT/NAACL*.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland, UK.

David Mimno, Hanna Wallach, Jason Naradowsky, David Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of EMNLP*.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the ACL/Coling*.

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

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Joseph Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates.* Ph.D. thesis, RWTH Aachen.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of ACL*.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of ACL*.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of ACL*.

Sadaf Abdul Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *Proceedings of EACL*.

Sujith Ravi and Kevin Knight. 2011. Deciphering foreign language. In *Proceedings of ACL/HLT*.

Vctor M. Sánchez-Cartagena, Felipe Sánchez-Martnez, and Juan Antonio Pérez-Ortiz. 2011. Integrating shallow-transfer rules into phrase-based statistical machine translation. In *Proceedings of the XIII Machine Translation Summit*.

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of CoNLL*.

Holger Schwenk and Jean Senellart. 2009. Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *MT Summit*.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of IWSLT*.

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of HLT/NAACL*.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT/NAACL*.

Christoph Tillmann. 2003. A projection extension algorithm for statistical machine translation. In *Proceedings of EMNLP*.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of CoLing*.

Ashish Venugopal, Stephan Vogel, and Alex Waibel. 2003. Effective phrase translation extraction from alignment models. In *Proceedings of ACL*.