Generative Models of Noisy Translations with Applications to Parallel Fragment Extraction

Chris Quirk*, **Raghavendra Udupa U.**†, **Arul Menezes***

*Microsoft Research
One Microsoft Way
Redmond, WA 98052
USA

†Microsoft Research India
196/36 2nd Main
Sadashivnagar, Banglaore 560 080
India
{chrisq,raghavu,arulm}@microsoft.com

Abstract

The development of broad domain statistical machine translation systems is gated by the availability of parallel data. A promising strategy for mitigating data scarcity is to mine parallel data from comparable corpora. Although comparable corpora seldom contain parallel sentences, they often contain parallel words or phrases. Recent fragment extraction approaches have shown that including parallel fragments in SMT training data can significantly improve translation quality. We describe efficient and effective generative models for extracting fragments, and demonstrate that these algorithms produce competitive improvements on cross-domain test data without suffering in-domain degradation even at very large scale.

1. Introduction

Statistical Machine Translation (SMT) systems are most influenced by two key components: language models and channel models. Language modeling still must overcome several obstacles (e.g. n-gram models are brittle especially with respect to morphology) but luckily data acquisition is not one of them – between large LDC corpora and the easy availability of web data, gigantic amounts of monolingual data are available. For channel modeling, however, the situation is much less promising. Most approaches require parallel data for training channel models, and derive continuing returns from larger datasets. Yet there are few large parallel corpora currently available. Even the largest (Arabic-English and Chinese-English) are orders of magnitude smaller than the available monolingual training data. These data sources also tend to be drawn from a single domain, and SMT systems trained on one domain suffer significant quality degradation when tested in other domains. If we hope to improve translation quality within a language pair and domain, expand to new domains, or acquire new language pairs, we must find ways to exploit non-parallel data sources.

1.1. Related work

There are many ways that we can identify and exploit comparable data. Finding comparable documents is a useful way point in this difficult task: we can significantly reduce the search space of further steps in the pipeline if we limit our attention to information in similar documents. Document pairs can be found from the web by exploiting URL structure, document structure, and lexical similarity amongst other clues (see for instance Resnik and Smith (2003), Zhang et al. (2006), Shi et al. (2006)). Alternatively we can search within large newswire corpora, which can be a rich source of translation information. Cross-lingual information retrieval techniques can find promising document pairs from large newswire corpora in different languages, as in (Zhao and Vogel, 2002). Although the web data is likely to be larger and more diverse, it presents obstacles to controlled experimentation (the web is constantly changing) and is seldom as carefully edited as newswire data. Therefore we focus on the latter source, but we expect that the techniques developed here should also apply to other sources of comparable data.

We loosely use the term “comparable” to describe the document pairs that can be extracted from these newswire sources, though the pairs differ significantly in translational equivalence. Occasionally the articles contain sentence-for-sentence translations of one another; there have been several efforts to search for whole-sentence translation pairs within comparable corpora (e.g., Zhao and Vogel (2002), Fung and Cheung (2004b), Fung and Cheung (2004a), Cheung and Fung (2004)). More often it appears that either two reporters have witnessed the same events and written similar accounts or perhaps one reporter has read another reporter’s account and subsequently written a new text with some common information. The latter articles contain few sentence-for-sentence translation pairs. Many researchers have instead tried to gather a bilingual lexicon from these sources, which could then be used by an MT system or even human translators (Fung and Yee (1998), Rapp (1999), Diab and Finch (2000), Koehn and Knight (1999), Gaussier et al. (2004), Shao and Ng (2004)). However comparable corpora contain multi-word translation information that is overlooked by these methods. For instance, quoted material from primary sources is often translated literally, as are person names, institution names, and other named entities.

We believe that one of the most promising ideas is to identify parallel sub-sentential fragments within comparable corpora, as proposed by Munteanu and Marcu (2006). Starting with a non-parallel corpus consisting of news articles from three sources (the BBC, the Romanian newspapers ‘Evenimentul Zilei’ and ‘Ziua’) they first produce a set of similar article pairs using a cross-lingual information retrieval system. Restricting their attention to sentence pairs that contain at least minimal lexical overlap, they search for parallel fragments using an approach inspired by signal processing. Using a set of parameters derived from LLR scores, they annotate each word with a value between -1 and 1 indicating the likelihood that this word has some translational equivalent in the other sentence by performing a greedy alignment. (Note that this step is performed without regard for position in the sentence, number of words generated by a single word, etc.) This stream of values is then treated as a signal and passed through a moving average filter. To find fragments—substrings of the original sentence that are likely to have a translation pair in the other side, they identify the longest spans that have only positive signal values. All fragments longer than some threshold (3 words) are concatenated to form a subsequence of the sentence that is likely to have a translation on the other side. The same process is repeated on the other sentence, and the resulting fragment pair is assumed to be parallel. They report substantial positive effect on BLEU scores when mined fragments are appended to a baseline parallel corpus.

We believe that there are several ways to improve this approach. Since the greedy alignment is performed independently in each direction, there is no guarantee that the words aligned to a fragment in one sentence will appear in the fragment from the other sentence. Nor is the number of spans guaranteed to match; the resulting spans are simply concatenated, which could produce odd phrases spanning fragment boundaries. The method does not model phenomena that have proven very important in the related task of word alignment, such as locality and fertility. Finally, the structure of the model is somewhat heuristic and thus difficult to optimize or chain in a pipelined process.

1.2. Our approach

We are primarily interested in the problem of extracting parallel fragments, particularly in developing theoretically-grounded, effective models. We present two algorithms for mining parallel fragments from similar sentence pairs, both based on generative models of semi-parallel data. First we train translational equivalence models from parallel data and monolingual generation models from source and target language data. Next, given two news wire corpora, we identify promising sentence pairs using methods very similar to those used in Munteanu and Marcu (2006). Our main innovation comes in identifying the parallel fragments from these comparable sources. In section 2., we describe two new models for extracting parallel fragments, and provide algorithms for effectively using them. We describe the experimental setup and empirical findings in section 3.. In section 4. we discuss our results and we propose some ideas for future exploration.

2. Generative models of fragment alignment

In most prior work (e.g. Brown et al. (1993), Vogel et al. (1996)), generative models are used to approximate the translation process. Given a sentence in one language (arbitrarily designed the *source*, denoted $\mathbf{s} = s_1^m$), we can find a probability distribution over sentences in the other language (designated *target*, denoted $\mathbf{t} = t_1^n$). While these models do allow for a certain degree of deviation between sentences, the deviations are assumed to be systematic (e.g. the Spanish word *de* must often be inserted when generating based on an English string). In noisy comparable sentences, the situation is markedly different: words may be inserted or deleted seemingly at random depending on what information each sentence happened to include. We describe two models to handle these phenomena: a conditional model of loose rather than exact translation, and a joint model of simultaneous generation.

2.1. Model A: Conditional generation

In the case of noisy translation, we assume that source language sentence has already been written, and the target language sentence is generated conditionally based on that sentence. Unlike standard word alignment models, however, we also allow words to be generated completely independently of the source language, based on prior target language words only. The intuition is as follows: say we already have a monolingual generation model (e.g., an n-gram language model), and a model of translational equivalence (e.g., an HMM word alignment model). We hypothesize that parallel fragments are more likely to be produced by translating the source sentence rather than a monolingual model: $\mathbf{Pr}(\mathbf{t}|\mathbf{s}) > \mathbf{Pr}(\mathbf{t})$.

Before describing the particulars of this model, we review some standard word alignment models. The generative framework behind IBM Models 1, 2, and the HMM model produces target language sentences in left-to-right order in the following manner. First the target sentence length is drawn according to an unspecified distribution – this detail is not important for the word alignment case. Next, for each target position, the position of the source word that generated this word is picked. Then the target word in that position is drawn according to the source word that generated that position. Let s_1^m be the source sentence, t_1^n be the target sentence, and $a_1^n \in \{0..m\}^n$ be the hidden state denoting the position of the source word generating each target word; $a_j = 0$ indicates that t_j was generated from the null word. Then we can model a sentence and an alignment as follows:

$$\begin{aligned}\mathbf{Pr}\left(a_1^n, t_1^n|s_1^m\right) = \mathbf{Pr}(|\mathbf{t}| = n)\cdot \\ \prod_{j=1}^{n}\big(\mathbf{Pr}(a_j|a_1^{j-1}, t_1^{j-1}, s_1^m)\cdot \\ \mathbf{Pr}(t_j|a_1^j, t_1^{j-1}, s_1^m)\big)\end{aligned}$$

All three models draw $\mathbf{Pr}(t_j|a_1^j, t_1^{j-1}, s_1^m)$ from $e(t_j|s_{a_j})$, a multinomial distribution conditioned on s_{a_j}. By drawing a_j from a uniform distribution, from a multinomial distribution based on j, or a multinomial distribution based on a_{j-1}, we produce IBM Model 1, Model 2, and the HMM model respectively.

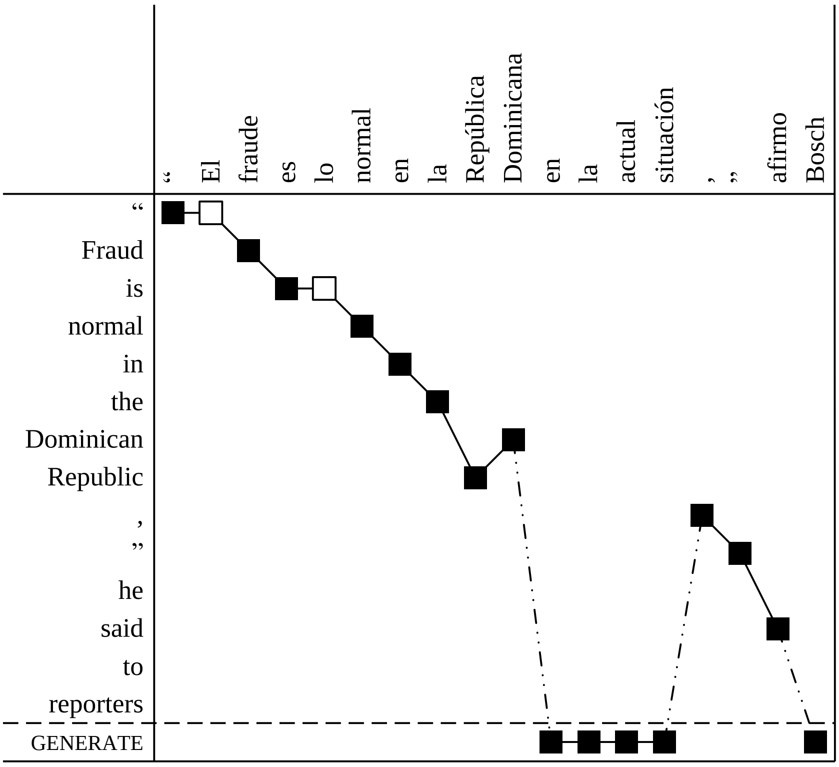

Figure 1: An example pair with a Model A alignment, demonstrating generation of the Spanish sentence given the English sentence. Each column contains a single box indicating the hidden state generating each Spanish word. In all but the last row, empty boxes represent null alignments and solid boxes represent actual alignments. The last row represents the hidden state indicating that the Spanish word was generated monolingually, independently of the English side.

To handle comparable data alignment, Model A augments the hidden space to include an additional state signifying that the current word is generated monolingually, without any corresponding source word. Note that this situation is distinct from the null word case: rather than drawing t_j according to source language information, we draw t_j according to previously generated target material. Let $a_j = -1$ indicate the monolingual generation state. Model A has the following structure for generating target words:

$$\mathbf{Pr}(t_j|a_1^j, t_1^{j-1}, s_1^m) = \begin{cases} e(t_j|t_1^{j-1}) & \text{if } a_j = -1 \\ e(t_j|s_{a_j}) & \text{otherwise} \end{cases}$$

and has a first-order dependence in the hidden states:

$$\mathbf{Pr}(a_j|a_1^{j-1}, t_1^{j-1}, s_1^m) = d(a_j|a_{j-1})$$

Parameter estimation. Model A parameters are defined in terms of standard HMM word alignment parameters and n-gram language model parameters. We begin by estimating a HMM parameters on a parallel corpus using EM; let the emission parameters be denoted with $e'(t|s)$ and the transition parameters be denoted $d'(a|a')$. In addition, we estimate a target language model based on monolingual data; the parameters be defined as: $e'(t_j|t_1^{j-1})$. Finally we are given free parameters φ indicating the probability of transitioning between bilingual BI and monolingual MO states.

Model A has in the same generative structure as the IBM models 1 and 2 and the HMM model, so it suffices to define how states and words are drawn. Transitioning between states is a straightforward mixture of φ and the HMM state model d', except that when jumping into a bilingual state, we consider all starting points equally likely:

$$\mathbf{Pr}(a_j|a_1^{j-1}, t_1^{j-1}, s_1^m) = \begin{cases} \varphi(\text{BI}|\text{BI})d'(a_j|a_1^{j-1}) & \text{if } a_{j-1} \neq -1, a_j \neq -1 \\ \varphi(\text{BI}|\text{MO})1/m & \text{if } a_{j-1} = -1, a_j \neq -1 \\ \varphi(\text{MO}|\text{BI}) & \text{if } a_{j-1} \neq -1, a_j = -1 \\ \varphi(\text{MO}|\text{MO}) & \text{if } a_{j-1} = -1, a_j = -1 \end{cases}$$

When generating words, we switch between monolingual and bilingual generation models according to the current hidden state.

$$\mathbf{Pr}(t_j|a_1^j, t_1^{j-1}, s_1^m) = \begin{cases} e'(t_j|t_1^{j-1}) & \text{if } a_j = -1 \\ e'(t_j|s_{a_j}) & \text{otherwise} \end{cases}$$

Fragment extraction. To find parallel fragments given a similar sentence pair, we first find the most likely hidden structure according to Model A. Since Model A has an HMM like structure, the familiar Viterbi algorithm will find the most likely alignment $\hat{\mathbf{a}} = \arg\max_{\mathbf{a}}\{\mathbf{Pr}(\mathbf{t}, \mathbf{a}|\mathbf{s})\}$.

Next we consider each maximal bilingual span (k, l); that is, for all $k' \in k..l$, $a_{k'} \neq -1$, and $a_{k-1} = a_{l+1} = -1$. Let i and j be the minimal and maximal non-zero values taken on by $\hat{\mathbf{a}}$ between k and l. Then we consider s_i^j and t_k^l to be a bilingual fragment pair if the following conditions hold:

1. Both fragments are at least the minimum length (currently 3).
2. The fraction of "holes" (unaligned words) in either span does not exceed a given threshold (30%).
3. The fraction of stop-words in either span does not exceed a given threshold (70%).

This model is attractive given that it is sound, relatively easy to implement, and quick to evaluate. However there are several aspects of the model which we might hope to improve. First, there are many free parameters to be tuned, including the transition probabilities φ and the threshold values. Second, the asymmetry of the model means that it does not forbid the same source fragment from generating multiple potentially overlapping fragments, nor does it evaluate the likelihood of the segmentation of the source side.

2.2. Model B: Joint generation

To address some of these limitations, we explore a joint model based on a slightly different generative decomposition. Rather that conditioning on one of the sentences, we generate the pair jointly. We imagine a process that chooses between three options: generate a source-only fragment, generate a target-only fragment, or generate a bilingual fragment in tandem. To further simplify the story, we can assume that the fragments are again generated left-to-right in both the source and target sentences. Although this assumption could potentially screen out non-monotone fragments, it also simplifies the model structure and search and is probably sufficient for language pairs with similar word order, such as English-Spanish.

The intuition behind Model B is similar to that of Model A, though phrased in joint rather than conditional probabilities. We hypothesize that the probability of generating source and target language fragments $\mathbf{s}$ and $\mathbf{t}$ jointly should be more likely than generating them independently (i.e. $\mathbf{Pr}(\mathbf{s}, \mathbf{t}) > \mathbf{Pr}(\mathbf{s}) \cdot \mathbf{Pr}(\mathbf{t})$) if and only if they are parallel. However few joint models of translation have proven effective in practice; conditional models have proven more effective empirically on most MT tasks. We can use Bayes' rule to combine a marginal probability and a conditional probability to estimate a joint probability: $\mathbf{Pr}(s, t) = \mathbf{Pr}(s)\mathbf{Pr}(t|s) = \mathbf{Pr}(t)|\mathbf{Pr}(t|s)$, though each direction is likely to give a different estimate of this joint probability. To optimize the precision of the extracted fragments, we use the minimum of either decomposition as the estimate of the joint probability.[1]

In this model, the hidden structure is a series of fragments $\mathbf{f}$. Each fragment f_i is a 2 tuple where $f_{i,1}$ and $f_{i,2}$ indicate the last source and target words covered by fragment f_i. The generative framework takes the following form:

$$\begin{aligned}\mathbf{Pr}(f_1^p, s_1^m, t_1^n) &= \mathbf{Pr}(|\mathbf{f}| = p) \cdot \\ &\prod_{i=1}^{p} \mathbf{Pr}(f_i | f_1^{i-1}, s_1^{f_{i-1,1}}, t_1^{f_{i-1,2}}) \cdot \\ &\quad \mathbf{Pr}(s_{f_{i-1,1}}^{f_{i,1}}, t_{f_{i-1,2}}^{f_{i,2}} | f_1^i, s_1^{f_{i-1,1}}, t_1^{f_{i-1},2})\end{aligned}$$

We first predict the number of fragments. Then for each fragment, we predict the number of source and target words generated by that fragment. Finally, we generate the source and target words in each fragment.

Model B makes several simplifying assumptions over this generative model. It assumes a uniform distribution of the number of fragments, and the number of source and target words generated by any one fragment. The final stage, the probability distribution over words within a fragment, is derived from independently estimated models. As in the conditional models, we use the HMM word alignment model for the conditional models with parameters estimated on given parallel corpus, and n-gram language models as marginal models with parameters estimated on monolingual corpora.

Parameter estimation. Let $L_s(\mathbf{s})$ and $L_t(\mathbf{t})$ be source and target n-gram language models, and $X_{st}(\mathbf{t}|\mathbf{s})$ and $X_{ts}(\mathbf{s}|\mathbf{t})$ be conditional translation models. We find the conditional probabilities X_{st}, X_{ts} by marginalizing over hidden alignments using the forward algorithm for HMM. Given these individual models, the Model B score of generating a fragment is:

$$\begin{aligned}&\mathbf{Pr}(s_{f_{i-1,1}}^{f_{i,1}}, t_{f_{i-1,2}}^{f_{i,2}} | f_1^i, s_1^{f_{i-1,1}}, t_1^{f_{i-1,2}}) = \min \\ &\quad \Bigg\{ \left(\prod_{j=f_{i-1,1}+1}^{f_{i,1}} L_s(s_j | s_1^{j-1}) \right) \cdot X_{st}(t_{f_{i-1,2}}^{f_{i,2}} | s_{f_{i-1,1}}^{f_{i,1}}), \\ &\quad \left(\prod_{k=f_{i-1,2}+1}^{f_{i,2}} L_t(t_k | t_1^{k-1}) \right) \cdot X_{ts}(s_{f_{i-1,1}}^{f_{i,1}} | t_{f_{i-1,2}}^{f_{i,2}}) \Bigg\}\end{aligned}$$

Fragment extraction. With Model B, we simply need to search for the most likely sequence of fragments $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \{\mathbf{Pr}(\mathbf{f}, \mathbf{s}, \mathbf{t})\}$. Note that each fragment need only condition on the endpoint of the previous fragment; in a sense this is like a first-order Markov model. This suggests a simple dynamic programming algorithm for finding the most likely alignment.

First we can precompute the cost of individual operations:

$$\begin{aligned}A[i,j] &= \prod_{x=i}^{j} L_s(s_x | s_1^{x-1}) \\ B[k,l] &= \prod_{x=k}^{l} L_t(t_x | t_1^{x-1}) \\ C[i,j,k,l] &= X_{ts}(s_i^j | t_k^l) \\ D[i,j,k,l] &= X_{st}(t_k^l | s_i^j) \\ E[i,j,k,l] &= \max\{A[i,j] \cdot D[i,j,k,l], \\ &\qquad B[k,l] \cdot C[i,j,k,l]\}\end{aligned}$$

[1] This is equivalent to saying that a fraction is considered parallel iff $\mathbf{Pr}(\mathbf{s}|\mathbf{t}) > \mathbf{Pr}(\mathbf{s})$ and $\mathbf{Pr}(\mathbf{t}|\mathbf{s}) > \mathbf{Pr}(\mathbf{t})$.

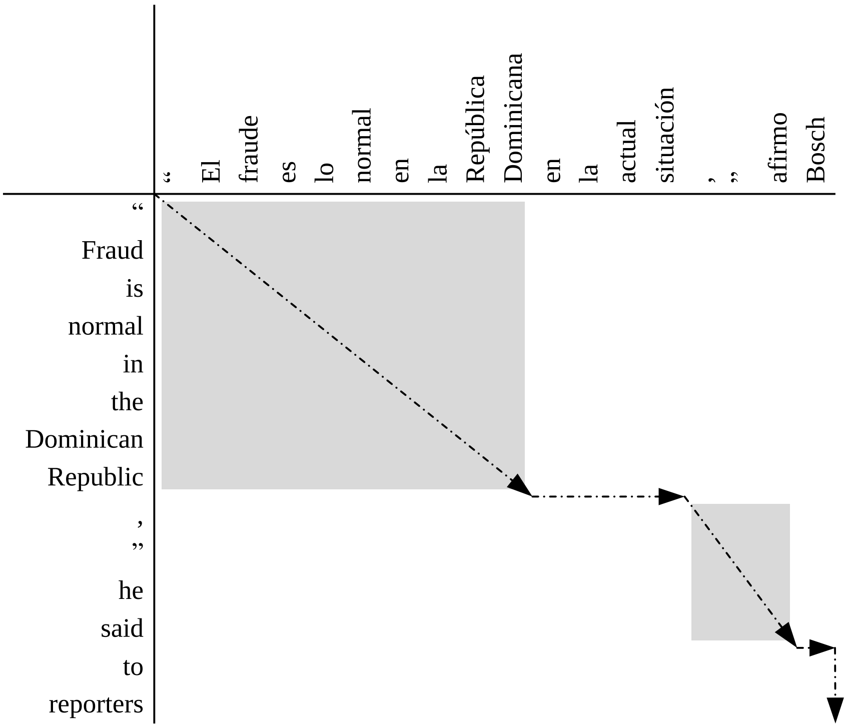

Figure 2: An example Model B alignment. This model jointly generates the sentence pair; arrows indicate the order in which each segment is generated. Since the Model B alignments are monotone, arrows may not point upwards or leftwards. Joint segments correspond to diagonal arrows; the parallel substring pairs are indicated by shaded areas.

Then the probability of the most likely fragment sequence ending at position (j, l) is

$$\begin{aligned} \delta[0,0] &= 1 \\ \delta[j,l] &= \max_{\substack{0 \le i \le j, \\ 0 \le k \le l}} \left\{ \begin{array}{l} \delta[i,l] \cdot A[i,j], \\ \delta[j,k] \cdot B[k,l], \\ \delta[i,k] \cdot E[i,j,k,l] \end{array} \right\} \end{aligned}$$

Unfortunately this algorithm is prohibitively expensive to evaluate, requiring runtime on the order of $O(m^2 n^3 + m^3 n^2)$ — even evaluating the E array is simply not feasible on tens or hundreds of millions of sentence pairs.

Many of the regions in this space are obviously not worth exploring. To help guide our search, we can compute some admissible heuristics:

$$\begin{aligned} C'[i,j] &= X_{ts}(s_i^j | \mathbf{t}) \\ D'[k,l] &= X_{st}(t_k^l | \mathbf{s}) \end{aligned}$$

When using the HMM model, we know that $C'[i,j] = X_{ts}(s_i^j|\mathbf{t}) \geq X_{ts}(s_i^j|t_k^l) = C[i,j,k,l]$ for any k, l since augmenting the state space only creates new paths and therefore adds total probability. Given this estimate, we can screen out any target span where $C'[i,j] < A[i,j]$ as they will never participate in the most likely fragment decomposition; the corresponding filter also applies to the target side.

Admissible future cost estimates can provide further reordering evidence:

$$\begin{aligned} F[i] &= \max_{i<j\le m} \{A[i,j] \cdot F[j], D'[i,j] \cdot F[j]\} \\ G[k] &= \max_{k<l\le n} \{B[k,l] \cdot G[l], C'[k,l] \cdot G[l]\} \end{aligned}$$

The cost of generating the remainder of the sentence pair from point (i, j) is bounded from above by $F[i] \cdot G[j]$. These future costs can be used in A* search or beam search. We found that keeping a beam of size 10 for each total number of words covered in both sentences leads to only a modest degradation in model cost while achieving orders of magnitude speedup over a true DP solution. Also limiting the maximum fragment size (to 12) and source to target fragment length ratios (to between 0.5 and 2) has a negligible impact on the fragments extracted while further increasing speed.

3. Experiments

To evaluate the efficacy of our methods, we focus on an extrinsic evaluation: the impact of the extracted parallel fragments on end-to-end machine translation quality. Intrinsic evaluation metrics could be quite useful for error analysis, optimization of free parameters, and faster experimental turnaround. However one must first demonstrate a correlation between such an intrinsic measure and the overall task accuracy; we therefore postpone this problem to future work.

	Spanish	English
Training sentences	730,740	
Words	15,725,136	15,222,505
Vocabulary	102,885	64,122
Dev sentences	2,000	
Words	60,628	58,655
Vocabulary	7,681	6,144
Devtest sentences	2,000	
Words	60,332	57,951
Vocabulary	7,782	6,054
Test sentences	3,064	
Words	91,730	85,232
Vocabulary	10,529	8,390

Table 1: Characteristics of the parallel seed corpus.

	Spanish	English
Documents	2,223,117	3,479,870
Sentences	20,177,725	49,293,904
Words	686,902,169	1,767,840,671
Low recall		
Document pairs	27,253,262	
Sentence pairs	2,660,283	
High recall		
Document pairs	27,985,397	
Sentence pairs	83,640,447	

Table 2: Characteristics of the first pass extraction; size of the raw Gigaword corpora as well as the selected document and sentence pair sets under two parameter settings.

3.1. Data sources

As a seed parallel corpus, we use the English-Spanish portion of the Europarl corpus (Koehn and Monz, 2006). Table 1 lists pertinent characteristics of this dataset. We train an HMM alignment model on this parallel data using GIZA++ (Och and Ney, 2003) using 5 iterations of model 1 followed by 5 iterations of the HMM model. This is performed symmetrically in both directions to produce conditional models of Spanish given English and English given Spanish. In addition to the Viterbi alignments, we save the HMM parameters for use in Models A and B. We also build trigram language models (smoothed using modified Kneser-Ney (Goodman, 2001)) on each side to be used both in decoding and in Models A and B.

The LDC English and Spanish Gigaword corpora (Graff, 2003; Graff, 2006) are a fertile ground for noisy word and phrase alignment. The articles draw from several major news feeds; both sides include articles from Agence France Presse, the Associated Press, and Xinhua News Agency, and the English side contains articles from the New York Times. While some article pairs may be rather close translations, more often it appears that stores are written mostly independently by English and Spanish authors. Therefore the articles are unlikely to contain parallel sentence pairs except in the case of direct quotes, which are often rendered exactly in one language and as a close translation in the other. On the other hand, named entities and other pairs do commonly occur in parallel.

3.2. First pass filtering

Invoking the fragment identification algorithm on all sentence pairs of the Gigaword corpus is obviously quite intractable. Therefore the first portion of our pipeline closely resembles Munteanu and Marcu (2006): promising document pairs are found using cross-lingual information retrieval techniques, then promising sentence pairs are found from amongst those documents. First we index the English Gigaword corpus. Next each Spanish article is translated into an English bag of words using the t-table from the HMM model: for all Spanish words s in the article, we append the English word e to the bag of words if the translation probability is above a threshold. This bag of words is issued as a query against the English corpus, and the top 20 English documents published within 7 days of the Spanish corpus are retrieved to form our promising document set.[2] While it is theoretically possible to run the noisy alignment algorithm on all sentence pairs in all promising documents, runtime is significantly diminished if we first filter the set down to a promising set of sentence pairs. To limit the sentences we consider, we require that the source sentence be neither more than twice as long as source, nor more than twice as short. Also we require that some fraction of the words in each sentence must have a translation in the other sentence.

By adjusting these constants we can trade off recall of potential translation information against the cleanliness of the retrieved data. We explore two settings:

1. A clean, high precision set with a t-table threshold of 0.125, where at least 5 or 40% of the words (whichever is greater) in each sentence must have a translation in the other side.

2. A noisier, high recall set with a t-table threshold of 0.1, with the minimum lowered to 2 words or 30% of the sentence.

Table 3 summarizes the results of this initial pass.

3.3. Fragment extraction

Next we apply several parallel fragment extraction approaches to these promising sentence pairs. As a baseline, we attempted to faithfully reimplement the approach described by Munteanu and Marcu (2006), which we will refer to as MM. We also applied Models A and B as described in the text. As Model A is a conditional model, the resulting fragments are asymmetric. We only evaluate it in one direction in this paper: predict English given Spanish.

The outputs from each algorithm are shown in Table 3. Yield from the MM algorithm is quite high; in fact, the yield from MM and Model B on the high recall data was so large that we could not easily employ it inside an MT system. Therefore we also evaluated each system on high recall data from only the year 2001, approximately one-sixth of the promising sentence pairs.

[2] We use BM25 (Robertson et al., 1995) ranking with the free parameter values $k_1 = 18$, $k_3 = 0.54$, and $b = 0.65$.

	Fragments	Spanish words	English words
Low recall			
MM	2,648,290	63,223,784	62,076,668
A($e\|s$)	1,529,801	8,526,713	8,656,013
B	2,993,201	24,342,112	25,222,313
High recall, 2001 only			
MM	3,780,631	64,193,250	68,096,131
A($e\|s$)	1,167,768	5,922,888	5,658,089
B	9,711,901	36,860,578	37,197,365
High recall, all years			
A($e\|s$)	6,626,777	33,382,123	31,827,777

Table 3: Yield from various fragment extraction algorithms.

The differences across algorithms are striking. The MM miner, for instance, has a very high yield. A large percentage of the sentences produce fragments, and often those fragments are on average quite long; unusually long in fact. It is difficult to evaluate precision and recall in these settings, but inspection suggests that many of these fragments are spurious.

Another unusual trend is that although Spanish sentences tend to be longer than English sentences, in several cases the models produce Spanish fragments that are shorter than their English fragments. This may suggest that modeling the length of each fragment could produce more equivalent data pairs.

3.4. MT evaluation

Finally we evaluated each of these fragment sets inside a machine translation system. We concatenated the fragment pairs with the original training data, retrained the word alignments on the augmented data again using five iterations of Model 1 and the HMM Model, then extracted phrase tables (Koehn et al., 2003). The language model and phrase tables are then used in a phrasal decoder that faithfully reimplements Pharaoh (Koehn, 2004) to translate the given test sets. Parameter weights are trained for each data set independently using minimum error rate training (Och, 2003) on BLEU (Papineni et al., 2002) using the provided development test set.

The test set consists of $2,000$ in-domain sentences drawn from held-out parliamentary data, as well as $1,064$ out-of domain sentences drawn from news commentary web sites. As we see in Table 4, the extracted fragments can positively influence translation quality in the out-of-domain news commentary data. Adding the MM fragments had a significantly negative impact on in-domain quality; we suspect that this is due to the noisy fragments produced by this approach. As the number of fragments in the training data increased, the quality of MM suffered even more. In contrast, when using fragments from Models A and B the system was able to achieve all the gains of that of the MM fragments without degradation of in-domain quality. The system also appears more robust to larger data, though disappointingly it does not achieve incremental returns.

		test	
	devtest	in	out
baseline	29.6	28.7	22.1
small data			
MM	28.5	27.7	22.4
Model A($e\|s$)	29.5	28.7	22.4
Model B	29.0	28.5	22.3
large data, 2001			
MM	26.4	26.1	20.0
Model A($e\|s$)	29.3	28.6	22.3
Model B	29.0	28.2	22.3
large data, all years			
Model A($e\|s$)	29.1	28.3	22.3

Table 4: BLEU scores of translation systems built with various training data.

4. Conclusions

We have presented a novel extension of the word alignment model to account for noisy translations, described how to leverage such a model to extract parallel fragments from comparable new corpora, and demonstrated the impact of these fragments on a machine translation system. The models are not challenging to implement, and provide a principled means of extracting information from sources containing some shard information.

However there are many potential improvements to explore. If we limit ourselves to the problem of extracting fragments from comparable articles, there are several points in this pipeline that could benefit from optimization. The information retrieval step in the middle has many free parameters (e.g., BM25 constants) and algorithmic variants (query translation, thresholds, etc.) that may have a major impact of final yield. If we selected these parameters to maximize recall or accuracy against a test set, it would likely lead to greater impact on new test sets. The free parameters within the extraction models should also be optimized on some task-based measure.

We have applied only single pass extraction. Instead we might bootstrap our models: retrain the noisy models, and re-extract new fragments. Such methods could further increase the vocabulary of the MT system. Although only a small addition may occur on the each iteration, the gradual increase in vocabulary aggregated across iterations might lead to significant differences.

Better models could also increase fragment yield. The simplifying assumptions in both models may well be a limitation. For instance, the uniform assumptions over fragment count and length could be replaced with learned models. Furthermore monotone alignments are not sufficient for many language pairs; an extension of the joint model toward ITG (Wu, 1997) could relieve the restriction, though at a cost of greater computational complexity. Despite the hard constraints on reordering imposed by ITG, Wu and Fung (2005) found that an ITG model of translational equivalence was very effective in identifying translation pairs. Even a model with limited reordering is likely to improve over a purely monotone baseline.

Noisy translation models also have potential applications beyond fragment extraction. Even so-called parallel corpora often contain loose and noisy translations. Models that allow for sub-sequences of the sentence to not align as well may lead to better alignment quality and better extracted phrase tables.

5. Acknowledgments

We thank A. Kumaran and Monojit Choudhury for their comments and Dragos Stefan Munteanu for answering several questions on the specifics of his work.

6. References

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

Percy Cheung and Pascale Fung. 2004. Sentence alignment in parallel, comparable, and quasicomparable corpora. In *Proceedings of LREC*, Lisbon, Portugal.

Mona Diab and Steve Finch. 2000. A statistical wordlevel translation model for comparable corpora. In *RIAO*, Paris, France.

Pascale Fung and Percy Cheung. 2004a. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of EMNLP*, Barcelona, Spain.

Pascale Fung and Percy Cheung. 2004b. Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpora. In *Proceedings of COLING*, Geneva, Switzerland.

Pascale Fung and Lo Yuen Yee. 1998. An ir approach for translating new words from nonparallel, comparable texts. In *Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics*, Montreal, Canada.

Eric Gaussier, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Herve Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of ACL*, Barcelona, Spain.

Joshua Goodman. 2001. A bit of progress in language modeling. Technical Report MSR-TR-2001-72, Microsoft Research.

David Graff. 2003. LDC2003T05: English gigaword corpus.

Dave Graff. 2006. LDC2006T12: Spanish gigaword corpus.

Philipp Koehn and Kevin Knight. 1999. Estimating word translation probabilities from unrelated monolingual corpora using the em algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, Austin, Texas, USA.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June. Association for Computational Linguistics.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL*, pages 127–133, Edmonton, Canada, May.

Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 115–124, Washington, USA, September.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics*, Sydney, Australia.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelpha, Pennsylvania, USA.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of ACL*, Maryland, USA.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Mike Gatford, and A. Payne. 1995. Okapi at trec-4. In *Proceedings of the Text Retrieval Conference*, pages 73–86.

Li Shao and Hwee Tou Ng. 2004. Mining new word translations from comparable corpora. In *Proceedings of COLING*, Geneva, Switzerland.

Lei Shi, Cheng Nie, Ming Zhou, and Jianfeng Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *Joint Proceedings of the Association for Computational Linguistics and the International Conference on Computational Linguistics*, Sydney, Australia.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING*, pages 836–741, Copenhagen, Denmark.

Dekai Wu and Pascale Fung. 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In *Proceedings of IJCNLP*, Edinborough, Scotland.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, September.

Y. Zhang, K. Wu, J. Gao, and Phile Vines. 2006. Automatic acquisition of chinese-english parallel corpus from the web. In *Proceedings of the European Conference on Information Retrieval*, Imperial College, London.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the IEEE International Conference on Data Mining*, Maebashi City, Japan.