Improving Machine Translation Performance by Exploiting Non-Parallel Corpora

Dragos Stefan Munteanu*
Information Sciences Institute
University of Southern California

Daniel Marcu*
Information Sciences Institute
University of Southern California

We present a novel method for discovering parallel sentences in comparable, non-parallel corpora. We train a maximum entropy classifier that, given a pair of sentences, can reliably determine whether or not they are translations of each other. Using this approach, we extract parallel data from large Chinese, Arabic, and English non-parallel newspaper corpora. We evaluate the quality of the extracted data by showing that it improves the performance of a state-of-the-art statistical machine translation system. We also show that a good-quality MT system can be built from scratch by starting with a very small parallel corpus (100,000 words) and exploiting a large non-parallel corpus. Thus, our method can be applied with great benefit to language pairs for which only scarce resources are available.

1. Introduction

Parallel texts—texts that are translations of each other—are an important resource in many NLP applications. They provide indispensable training data for statistical machine translation (Brown et al. 1990; Och and Ney 2002) and have been found useful in research on automatic lexical acquisition (Gale and Church 1991; Melamed 1997), cross-language information retrieval (Davis and Dunning 1995; Oard 1997), and annotation projection (Diab and Resnik 2002; Yarowsky and Ngai 2001; Yarowsky, Ngai, and Wicentowski 2001).

Unfortunately, parallel texts are also scarce resources: limited in size, language coverage, and language register. There are relatively few language pairs for which parallel corpora of reasonable sizes are available; and even for those pairs, the corpora come mostly from one domain, that of political discourse (proceedings of the Canadian or European Parliament, or of the United Nations). This is especially problematic for the field of statistical machine translation (SMT), because translation systems trained on data from a particular domain (e.g., parliamentary proceedings) will perform poorly when translating texts from a different domain (e.g., news articles).

One way to alleviate this lack of parallel data is to exploit a much more available and diverse resource: comparable non-parallel corpora. Comparable corpora are texts that, while not parallel in the strict sense, are somewhat related and convey overlapping information. Good examples are the multilingual news feeds produced by news agencies such as Agence France Presse, Xinhua News, Reuters, CNN, BBC, etc. Such texts are widely available on the Web for many language pairs and domains. They often

* 4676 Admiralty Way, Suite 1001, Marina del Rey, CA 90292. E-mail: {dragos,marcu}@isi.edu.

Submission received: 5 November 2004; Accepted for publication: 3 March 2005.

contain many sentence pairs that are fairly good translations of each other. The ability to reliably identify these pairs would enable the automatic creation of large and diverse parallel corpora.

However, identifying good translations in comparable corpora is hard. Even texts that convey the same information will exhibit great differences at the sentence level. Consider the two newspaper articles in Figure 1. They have been published by the English and French editors of Agence France Presse, and report on the same event, an epidemic of cholera in Pyongyang. The lines in the figure connect sentence pairs that are approximate translations of each other. Discovering these links automatically is clearly non-trivial. Traditional sentence alignment algorithms (Gale and Church 1991; Wu 1994; Fung and Church 1994; Melamed 1999; Moore 2002) are designed to align sentences in parallel corpora and operate on the assumption that there are no reorderings and only limited insertions and deletions between the two renderings of a parallel document. Thus, they perform poorly on comparable, non-parallel texts. What we need are methods able to judge sentence pairs in isolation, independent of the (potentially misleading) context.

This article describes a method for identifying parallel sentences in comparable corpora and builds on our earlier work on parallel sentence extraction (Munteanu, Fraser, and Marcu 2004). We describe how to build a maximum entropy-based classifier that can reliably judge whether two sentences are translations of each other, without making use of any context. Using this classifier, we extract parallel sentences from very large comparable corpora of newspaper articles. We demonstrate the quality of our

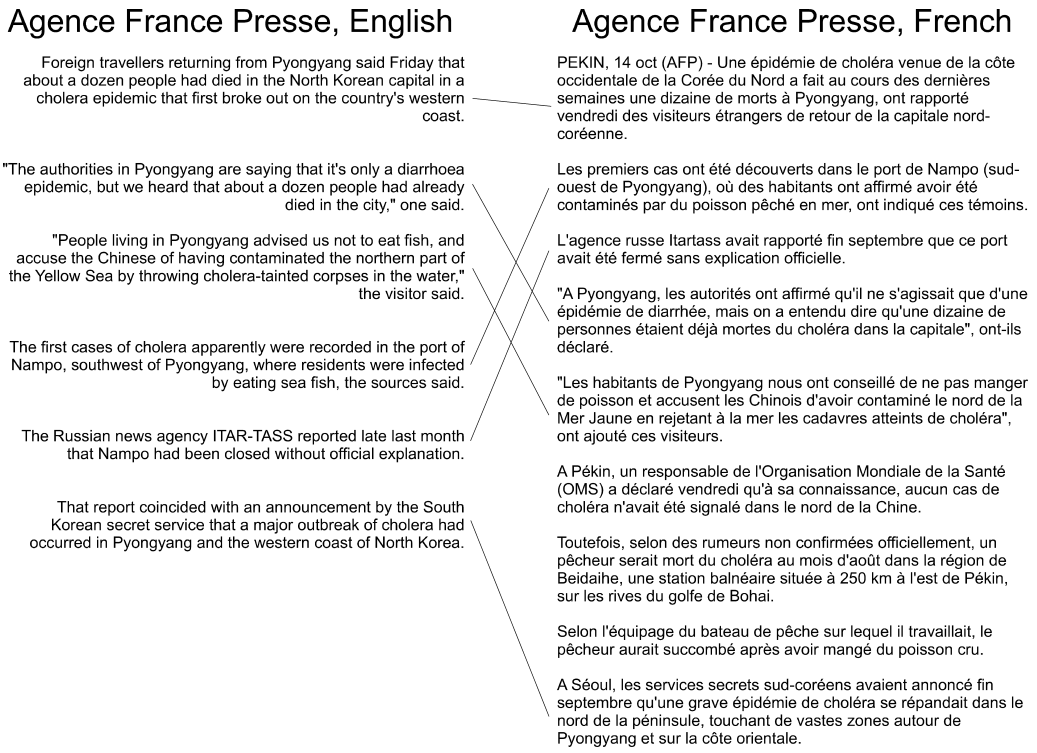


Figure 1
A pair of comparable texts.

extracted sentences by showing that adding them to the training data of an SMT system improves the system's performance. We also show that language pairs for which very little parallel data is available are likely to benefit the most from our method; by running our extraction system on a large comparable corpus in a bootstrapping manner, we can obtain performance improvements of more than 50% over a baseline MT system trained only on existing parallel data.

Our main experimental framework is designed to address the commonly encountered situation that exists when the MT training and test data come from different domains. In such a situation, the test data is *in-domain*, and the training data is *out-of-domain*. The problem is that in such conditions, translation performance is quite poor; the out-of-domain data doesn't really help the system to produce good translations. What is needed is additional in-domain training data. Our goal is to get such data from a large in-domain comparable corpus and use it to improve the performance of an out-of-domain MT system. We work in the context of Arabic-English and Chinese-English statistical machine translation systems. Our *out-of-domain* data comes from translated United Nations proceedings, and our *in-domain* data consists of news articles. In this experimental framework we have access to a variety of resources, all of which are available from the Linguistic Data Consortium:[1]

- large amounts of *out-of-domain* parallel data;
- smaller amounts of *in-domain* parallel data;
- *in-domain* MT test corpora with four reference translations; and
- *in-domain* comparable corpora: large collections of Arabic, Chinese, and English news articles from various news agencies.

In summary, we call *in-domain* the domain of the test data that we wish to translate; in this article, that in-domain data consists of news articles. *Out-of-domain* data is data that belongs to any other domain; in this article, the out-of-domain data is drawn from United Nations (UN) parliamentary proceedings. We are interested in the situation that exists when we need to translate news data but only have UN data available for training. The solution we propose is to get comparable news data, automatically extract parallel sentences from it, and use these sentences as additional training data; we will show that doing this improves translation performance on a news test set. The Arabic-English and Chinese-English resources described in the previous paragraph enable us to simulate our conditions of interest and perform detailed measurements of the impact of our proposed solution. We can train baseline systems on UN parallel data (using the data from the first bullet in the previous paragraph), extract additional news data from the large comparable corpora (the fourth bullet), accurately measure translation performance on news data against four reference translations (the third bullet), and compare the impact of the automatically extracted news data with that of similar amounts of human-translated news data (the second bullet).

In the next section, we give a high-level overview of our parallel sentence extraction system. In Section 3, we describe in detail the core of the system, the parallel sen-

1 http://www.ldc.upenn.edu.

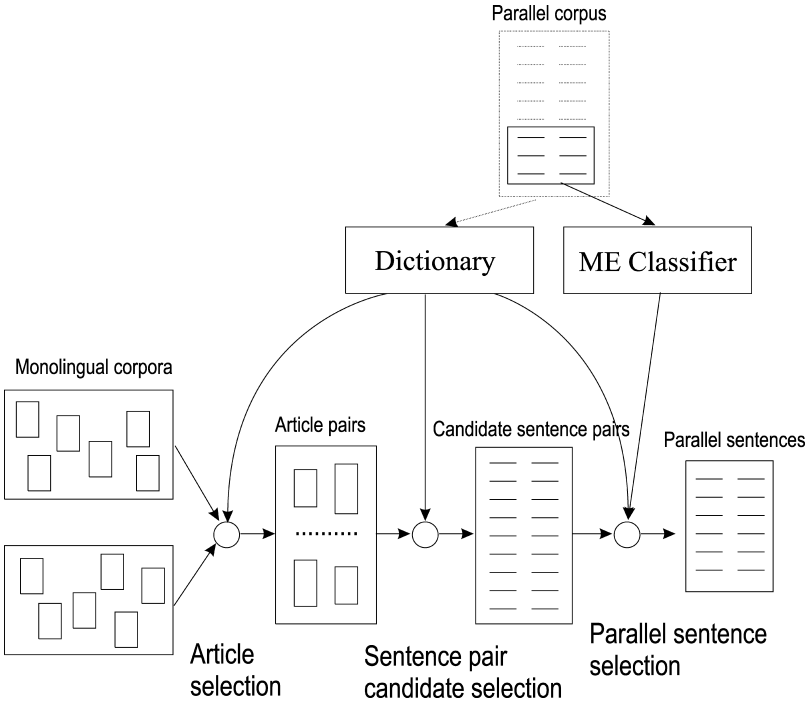


Figure 2
A Parallel Sentence Extraction System.

tence classifier. In Section 4, we discuss several data extraction experiments. In Section 5, we evaluate the extracted data by showing that adding it to out-of-domain parallel data improves the in-domain performance of an out-of-domain MT system, and in Section 6, we show that in certain cases, even larger improvements can be obtained by using bootstrapping. In Section 7, we present examples of sentence pairs extracted by our method and discuss some of its weaknesses. Before concluding, we discuss related work.

2. A System for Extracting Parallel Sentences from Comparable Corpora

The general architecture of our extraction system is presented in Figure 2. Starting with two large monolingual corpora (a non-parallel corpus) divided into documents, we begin by selecting pairs of similar documents (Section 2.1). From each such pair, we generate all possible sentence pairs and pass them through a simple word-overlap-based filter (Section 2.2), thus obtaining candidate sentence pairs. The candidates are presented to a maximum entropy (ME) classifier (Section 2.3) that decides whether the sentences in each pair are mutual translations of each other.

The resources required by the system are minimal: a bilingual dictionary and a small amount of parallel data (used for training the ME classifier). The dictionaries used in our experiments are learned automatically from (out-of-domain) parallel corpora;[2] thus, the only resource used by our system consists of parallel sentences.

2 If such a resource is unavailable, other dictionaries can be used.

2.1 Article Selection

Our comparable corpus consists of two large, non-parallel, news corpora, one in English and the other in the foreign language of interest (in our case, Chinese or Arabic). The parallel sentence extraction process begins by selecting, for each foreign article, English articles that are likely to contain sentences that are parallel to those in the foreign one.

This step of the process emphasizes recall rather than precision. For each foreign document, we do not attempt to find the best-matching English document, but rather a set of similar English documents. The subsequent components of the system are robust enough to filter out the extra noise introduced by the selection of additional (possibly bad) English documents.

We perform document selection using the Lemur IR toolkit[3] (Ogilvie and Callan 2001). We first index all the English documents into a database. For each foreign document, we take the top five translations of each of its words (according to our probabilistic dictionary) and create an English language query. The translation probabilities are only used to choose the word translations; they do not appear in the query. We use the query to run TF-IDF retrieval against the database, take the top 20 English documents returned by Lemur, and pair each of them with the foreign query document.

This document matching procedure is both slow (it looks at all possible document pairs, so it is quadratic in the number of documents) and imprecise (due to noise in the dictionary, the query will contain many wrong words). We attempt to fix these problems by using the following heuristic: we consider it likely that articles with similar content have publication dates that are close to each other. Thus, each query is actually run only against English documents published within a window of five days around the publication date of the foreign query document; we retrieve the best 20 of these documents. Each query is thus run against fewer documents, so it becomes faster and has a better chance of getting the right documents at the top.

Our experiments have shown that the final performance of the system does not depend too much on the size of the window (for example, doubling the size to 10 days made no difference). However, having no window at all leads to a decrease in the overall performance of the system.

2.2 Candidate Sentence Pair Selection

From each foreign document and set of associated English documents, we take all possible sentence pairs and pass them through a word-overlap filter.

The filter verifies that the ratio of the lengths of the two sentences is no greater than two. It then checks that at least half the words in each sentence have a translation in the other sentence, according to the dictionary. Pairs that do not fulfill these two conditions are discarded. The others are passed on to the parallel sentence selection stage.

This step removes most of the noise (i.e., pairs of non-parallel sentences) introduced by our recall-oriented document selection procedure. It also removes good pairs that fail to pass the filter because the dictionary does not contain the necessary entries; but those pairs could not have been handled reliably anyway, so the overall effect of the filter is to improve the precision and robustness of the system. However, the filter also accepts many wrong pairs, because the word-overlap condition is weak; for instance, stopwords almost always have a translation on the other side, so if a few of the content

3 http://www-2.cs.cmu.edu/~lemur.

words happen to match, the overlap threshold is fulfilled and an erroneous candidate sentence pair is selected.

2.3 Parallel Sentence Selection

For each candidate sentence pair, we need a reliable way of deciding whether the two sentences in the pair are mutual translations. This is achieved by a Maximum Entropy (ME) classifier (described at length in Section 3), which is the core component of our system. Those pairs that are classified as being translations of each other constitute the output of the system.

3. A Maximum Entropy Classifier for Parallel Sentence Identification

In the Maximum Entropy (ME) statistical modeling framework, we impose constraints on the model of our data by defining a set of feature functions. These feature functions emphasize properties of the data that we believe to be useful for the modeling task. For example, for a sentence pair sp, the word overlap (the percentage of words in either sentence that have a translation in the other) might be a useful indicator of whether the sentences are parallel. We therefore define a feature function $f(sp)$, whose value is the word overlap of the sentences in sp.

According to the ME principle, the optimal parametric form of the model of our data, taking into account the constraints imposed by the feature functions, is a log linear combination of these functions. Thus, for our classification problem, we have:

$$P(c_i|sp) = \frac{1}{Z(sp)} \prod_{j=1}^{k} \lambda_j^{f_{ij}(c,sp)}$$

where c_i is the class (c_0="parallel", c_1="not parallel"), $Z(sp)$ is a normalization factor, and f_{ij} are the feature functions (indexed both by class and by feature). The resulting model has free parameters λ_j, the feature weights. The parameter values that maximize the likelihood of a given training corpus can be computed using various optimization algorithms (see [Malouf 2002] for a comparison of such algorithms).

3.1 Features for Parallel Sentence Identification

For our particular classification problem, we need to find feature functions that distinguish between parallel and non-parallel sentence pairs. For this purpose, we compute and exploit word-level alignments between the sentences in each pair. A word alignment between two sentences in different languages specifies which words in one sentence are translations of which words in the other. Word alignments were first introduced in the context of statistical MT, where they are used to estimate the parameters of a translation model (Brown et al. 1990). Since then, they were found useful in many other NLP applications (e.g., word sense tagging [Diab and Resnik 2002] and question answering [Echihabi and Marcu 2003]).

Figures 3 and 4 give examples of word alignments between two English-Arabic sentence pairs from our comparable corpus. Each figure contains two alignments. The one on the left is a correct alignment, produced by a human, while the one on the right

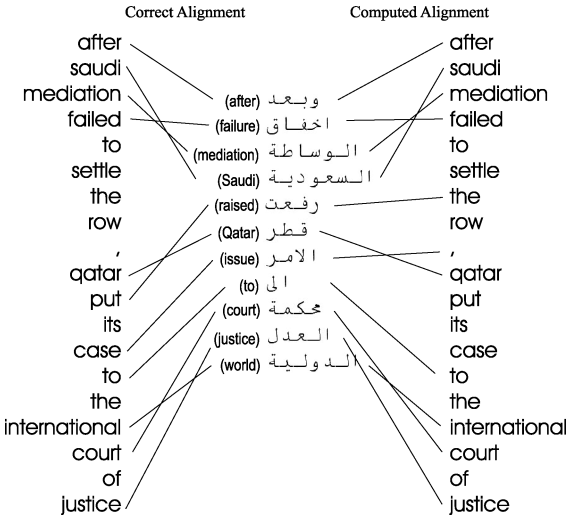


Figure 3
Alignments between two parallel sentences.

was computed automatically. As can be seen from the gloss next to the Arabic words, the sentences in Figure 3 are parallel while the sentences in Figure 4 are not.

In a correct alignment between two non-parallel sentences, most words would have no translation equivalents; in contrast, in an alignment between parallel sentences, most words would be aligned. Automatically computed alignments, however, may have incorrect connections; for example, on the right side of Figure 3, the Arabic word *issue* is connected to the comma; and in Figure 4, the Arabic word *at* is connected to the English phrase *its case to the*. Such errors are due to noisy dictionary entries and to

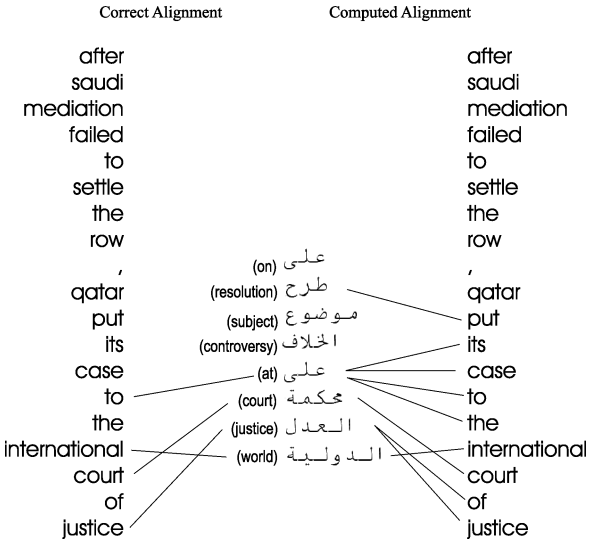


Figure 4
Alignments between two non-parallel sentences.

shortcomings of the model used to generate the alignments. Thus, merely looking at the number of unconnected words, while helpful, is not discriminative enough. Still, automatically produced alignments have certain additional characteristics that can be exploited.

We follow Brown et al. (1993) in defining the **fertility** of a word in an alignment as the number of words it is connected to. The presence, in an automatically computed alignment between a pair of sentences, of words of high fertility (such as the Arabic word *at* in Figure 4) is indicative of non-parallelism. Most likely, these connections were produced because of a lack of better alternatives.

Another aspect of interest is the presence of long **contiguous connected spans**, which we define as pairs of bilingual substrings in which the words in one substring are connected only to words in the other substring. Such a span may contain a few words without any connection (a small percentage of the length of the span), but no word with a connection outside the span. Examples of such spans can be seen in Figure 3: the English strings *after saudi mediation failed* or *to the international court of justice* together with their Arabic counterparts. Long contiguous connected spans are indicative of parallelism, since they suggest that the two sentences have long phrases in common. And, in contrast, long substrings whose words are all unconnected are indicative of non-parallelism.

To summarize, our classifier uses the following features, defined over two sentences and an automatically computed alignment between them.

General features (independent of the word alignment):

- lengths of the sentences, as well as the length difference and length ratio;
- percentage of words on each side that have a translation on the other side (according to the dictionary).

Alignment features:

- percentage and number of words that have no connection;
- the top three largest fertilities;
- length of the longest contiguous connected span; and
- length of the longest unconnected substring.

3.2 Word Alignment Model

In order to compute word alignments we need a simple and efficient model. We want to align a large number of sentences, with many out-of-vocabulary words, in reasonable time. We also want a model with as few parameters as possible—preferably only word-for-word translation probabilities.

One such model is the IBM Model 1 (Brown et al. 1993). According to this model, given foreign sentence ($f_{j1<=j<=m}$), English sentence ($e_{i1<=i<=l}$), and translation probabilities $t(f_j|e_i)$, the best alignment $f \rightarrow e$ is obtained by linking each foreign word f_j to its most likely English translation $argmax_{e_i} t(f_j|e_i)$. Thus, each foreign word is aligned to exactly one English word (or to a special NULL token).

Due to its simplicity, this model has several shortcomings, some more structural than others (see Moore [2004] for a discussion). Thus, we use a version that is augmented with two simple heuristics that attempt to alleviate some of these shortcomings.

One possible improvement concerns English words that appear more than once in a sentence. According to the model, a foreign word that prefers to be aligned with such an English word could be equally well aligned with any instance of that word. In such situations, instead of arbitrarily choosing the first instance or a random instance, we attempt to make a "smarter" decision. First, we create links only for those English words that appear exactly once; next, for words that appear more than once, we choose which instance to link with so that we minimize the number of crossings with already existing links.

The second heuristic attempts to improve the choice of the most likely English translation of a foreign word. Our translation probabilities are automatically learned from parallel data, and we learn values for both $t(f_j|e_i)$ and $t(e_i|f_j)$. We can therefore decide that the most likely English translation of f_j is $argmax_{e_i}\{t(f_j|e_i), t(e_i|f_j)\}$. Using both sets of probabilities is likely to help us make a better-informed decision.

Using this alignment strategy, we follow (Och and Ney 2003) and compute one alignment for each translation direction ($f \rightarrow e$ and $e \rightarrow f$), and then combine them. Och and Ney present three combination methods: *intersection*, *union*, and *refined* (a form of intersection expanded with certain additional neighboring links).

Thus, for each sentence pair, we compute five alignments (two modified-IBM-Model-1 plus three combinations) and then extract one set of general features and five sets of alignment features (as described in the previous section).

3.3 Training and Testing

We create training instances for our classifier from a small parallel corpus. The simplest way to obtain classifier training data from a parallel corpus is to generate all possible sentence pairs from the corpus (the Cartesian product). This generates $5{,}000^2$ training instances, out of which 5,000 are positive (i.e., belong to class "parallel") and the rest are negative.

One drawback of this approach is that the resulting training set is very imbalanced, i.e., it has many more negative examples than positive ones. Classifiers trained on such data do not achieve good performance; they generally tend to predict the majority class, i.e., classify most sentences as non-parallel (which has indeed been the case in our experiments). Our solution to this is to downsample, i.e., eliminate a number of (randomly selected) negative instances.

Another problem is that the large majority of sentence pairs in the Cartesian product have low word overlap (i.e., few words that are translations of each other). As explained in Section 2 (and shown in Figure 2), when extracting data from a comparable corpus, we only apply the classifier on the output of the word-overlap filter. Thus, low-overlap sentence pairs, which would be discarded by the filter, are unlikely to be useful as training examples. We therefore use for training only those pairs from the Cartesian product that are accepted by the word-overlap filter. This has the additional advantage that, since all these pairs have many words in common, the classifier learns to make distinctions that cannot be made based on word overlap alone.

To summarize, we prepare our classifier training set in the following manner: starting from a parallel corpus of about 5,000 sentence pairs, we generate all the sentence pairs in the Cartesian product; we discard the pairs that do not fulfill the conditions of the word-overlap filter; if the resulting set is imbalanced, i.e., the ratio of non-parallel to parallel pairs is greater than five, we balance it by removing randomly chosen non-parallel pairs. We then compute word alignments and extract feature values.

Using the training set, we compute values for the classifier feature weights using the YASMET[4] implementation of the GIS algorithm (Darroch and Ratcliff 1974). Since we are dealing with few parameters and have sufficiently many training instances, using more advanced training algorithms is unlikely to bring significant improvements.

We test the performance of the classifier by generating test instances from a different parallel corpus (also around 5,000 sentence pairs) and checking how many of these instances are correctly classified. We prepare the test set by creating the Cartesian product of the sentences in the test parallel corpus and applying the word-overlap filter (we do not perform any balancing). Although we apply the filter, we still conceptually classify all pairs from the Cartesian product in a two-stage classification process: all pairs discarded by the filter are classified as "non-parallel," and for the rest, we obtain predictions from the classifier. Since this is how we apply the system on truly unseen data, this is the process in whose performance we are interested.

We measure the performance of the classification process by computing **precision** and **recall. Precision** is the ratio of sentence pairs correctly judged as parallel to the total number of pairs judged as parallel by the classifier. **Recall** is the ratio of sentence pairs correctly identified as parallel by the classifier to the total number of truly parallel pairs—i.e., the number of pairs in the parallel corpus used to generate the test instances. Both numbers are expressed as percentages. More formally: let *classified_parallel* be the total number of sentence pairs from our test set that the classifier judged as parallel, *classified_well* be the number of pairs that the classifier correctly judged as parallel, and *true_parallel* be the total number of parallel pairs in the test set. Then:

$$precision = 100 * \frac{classified\_well}{classified\_parallel} \quad recall = 100 * \frac{classified\_well}{true\_parallel}$$

3.4 Performance Evaluation

There are two factors that influence a classifier's performance: dictionary coverage and similarity between the domains of the training and test instances. We performed evaluation experiments to account for both these factors.

All our dictionaries are automatically learned from parallel data; thus, we can create dictionaries of various coverage by learning them from parallel corpora of different sizes. We use five dictionaries, learned from five *initial* out-of-domain parallel corpora, whose sizes are 100k, 1M, 10M, 50M, and 95M tokens, as measured on the English side.

Since we want to use the classifier to extract sentence pairs from our in-domain comparable corpus, we test it on instances generated from an in-domain parallel corpus. In order to measure the effect of the domain difference, we use two training sets: one generated from an in-domain parallel corpus and another one from an out-of-domain parallel corpus.

In summary, for each language pair, we use the following corpora:

- five *initial* out-of-domain corpora of various sizes, used for learning dictionaries;
- one out-of-domain classifier training corpus;

4 http://www.fjoch.com/YASMET.html.

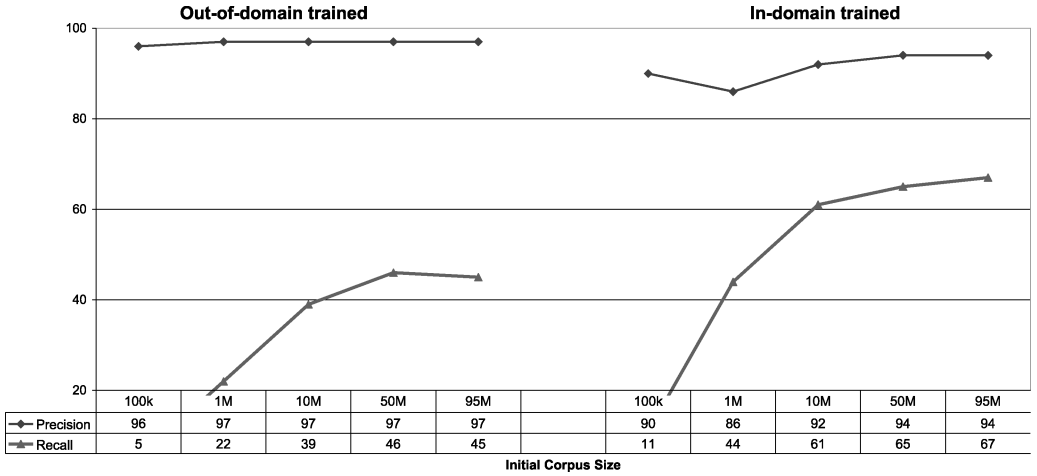


Figure 5
Precision and recall of the Arabic-English classifiers.

- one in-domain classifier training corpus; and
- one in-domain classifier test corpus.

From each initial, out-of-domain corpus, we learn a dictionary. We then take the classifier training and test corpora and, using the method described in the previous section, create two sets of training instances and one set of test instances. We train two classifiers (one on each training set) and evaluate both of them on the test set.

The parallel corpora used for generating training and test instances have around 5k sentence pairs each (approximately 150k English tokens), and generate around 10k training instances (for each training set) and 8k test instances.

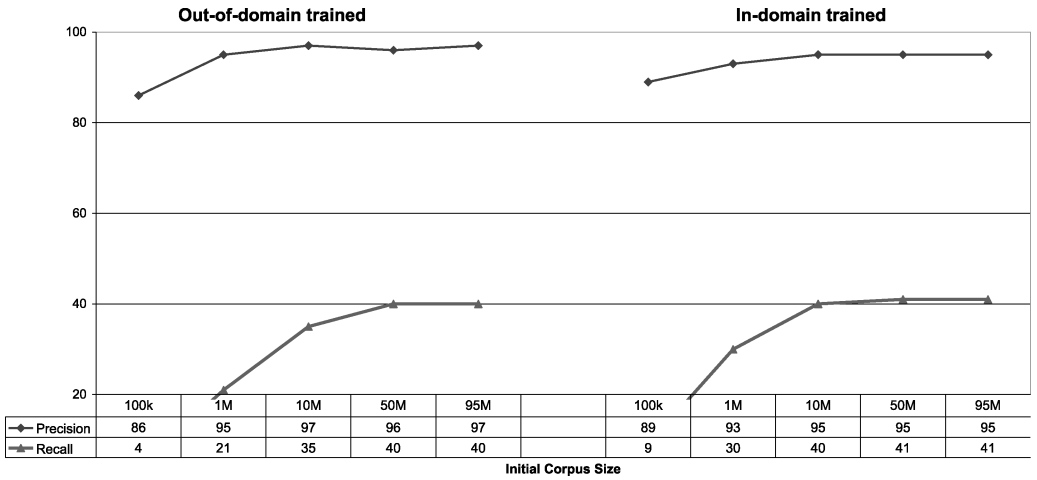


Figure 6
Precision and recall of the Chinese-English classifiers.

Figures 5 and 6 show the recall and precision of our classifiers, for both Arabic-English and Chinese-English. The results show that the precision of our classification process is robust with respect to dictionary coverage and training domain. Even when starting from a very small initial parallel corpus, we can build a high-precision classifier. Having a good dictionary and training data from the right domain does help though, mainly with respect to recall.

The classifiers achieve high precision because their positive training examples are clean parallel sentence pairs, with high word overlap (since the pairs with low overlap are filtered out); thus, the classification decision frontier is pushed towards "good-looking" alignments. The low recall results are partly due to the word-overlap filter (the first stage of the classification process), which discards many parallel pairs. If we don't apply the filter before the classifier, the recall results increase by about 20% (with no loss in precision). However, the filter plays a very important role in keeping the extraction pipeline robust and efficient (as shown in Figure 7, the filter discards 99% of the candidate pairs), so this loss of recall is a price worth paying.

Classifier evaluations using different subsets of features show that most of the classifier performance comes from the general features together with the alignment features concerning the percentage and number of words that have no connection. However, we expect that in real data, the differences between parallel and non-parallel pairs are less clear than in our test data (see the discussion in Section 7) and can no

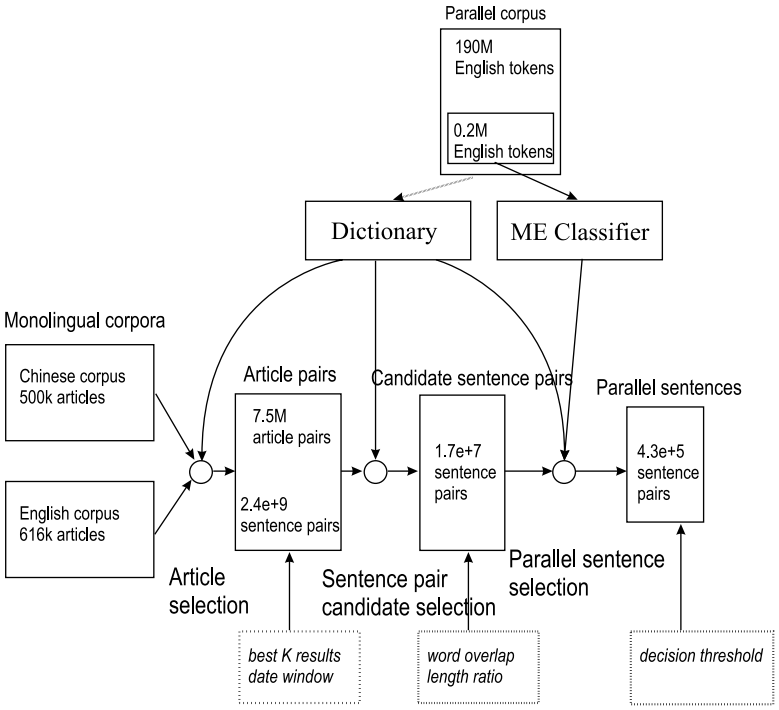


Figure 7
The amounts of data processed by our system during extraction from the Chinese-English comparable corpus.

Table 1
The Gigaword comparable corpora.

		Foreign		English	
Language pair	News agency and period	# articles	# tokens	# articles	# tokens
Arabic-English	AFP, 1994–1997, 2002 Xinhua News, 2001	224k	40M	650k	195M
Chinese-English	Xinhua News, 1995–2001	457k	162M	580k	128M

longer be accounted for only by counting the linked words; thus, the other features should become more important.

4. Data Extraction Experiments

4.1 Controlled Experiments

The comparable corpora that we use for parallel sentence extraction are collections of news stories published by the Agence France Presse and Xinhua News agencies. They are parts of the Arabic, English, and Chinese Gigaword corpora which are available from the Linguistic Data Consortium. From these collections, for each language pair, we create an in-domain comparable corpus by putting together articles coming from the same agency and the same time period. Table 1 presents in detail the sources and sizes of the resulting comparable corpora. The remainder of the section presents the various data sets that we extracted automatically from these corpora, under various experimental conditions.

In the experiments described in Section 3.4, we started out with five out-of-domain *initial* parallel corpora of various sizes and obtained five dictionaries and five out-of-domain trained classifiers (per language pair). We now plug in each of these classifiers (and their associated dictionaries) in our extraction system (Section 2) and apply it to our comparable corpora. We thus obtain five Arabic-English and five Chinese-English *extracted* corpora.

Note that in each of these experiments the only resource used by our system is the initial, out-of-domain parallel corpus. Thus, the experiments fit in the framework of interest described in Section 1, which assumes the availability of (limited amounts of) out-of-domain training data and (large amounts of) in-domain comparable data.

Table 2 shows the sizes of the extracted corpora for each *initial* corpus size, for both Chinese-English and Arabic-English. As can be seen, when the initial parallel corpus is very small, the amount of extracted data is also quite small. This is due to the low coverage of the dictionary learned from that corpus. Our candidate pair selection step (Section 2.2) discards pairs with too many unknown (or unrelated) words, according to the dictionary; thus, only few sentences fulfill the word-overlap condition of our filter.

As mentioned in Section 1, our goal is to use the extracted data as additional MT training data and obtain better translation performance on a given in-domain MT test set. A simple way of estimating the usefulness of the data for this purpose is to measure its *coverage* of the test set, i.e., the percentage of running *n*-grams from the test corpus that are also in our corpus. Tables 3 and 4 present the coverage of our

Table 2
Size of the datasets extracted from the comparable corpora, in millions of English words.

	Size of automatically extracted corpora	
Size of *initial* parallel corpus	Arabic-English	Chinese-English
100k	0.09M	0.9M
1M	0.6M	5M
10M	1.9M	8.3M
50M	2.2M	10.5M
95M	2.1M	10.5M

Table 3
Coverage of the extracted corpora for Arabic-English.

	Out-of-domain	In-domain
Initial corpus size	Initial	Initial plus extracted
100k	68/16/3/0.5	82/31/8/2
1M	86/33/7/1	94/54/20/7
10M	95/51/16/3	98/67/30/12
50M	98/64/24/6	99/74/36/14
95M	98/68/28/8	99/76/38/15

Table 4
Coverage of the extracted corpora for Chinese-English.

	Out-of-domain	In-domain
Initial corpus size	Initial	Initial plus extracted
100k	75/19/2/0.2	91/41/11/3
1M	90/38/8/1	97/61/22/7
10M	97/57/18/4	99/70/29/10
50M	98/69/27/7	99/76/36/12
95M	99/73/32/9	99/78/39/14

extracted corpora. For each *initial* corpus size, the first column shows the coverage of that initial corpus, and the second column shows the coverage of the initial corpus plus the extracted corpus. Each cell contains four numbers that represent the coverage with respect to unigrams, bigrams, trigrams, and 4-grams. The numbers show that unigram coverage depends only on the size of the corpus (and not on the domain), but for longer *n*-grams, our in-domain extracted data brings significant improvements in coverage.

4.2 Non-Controlled Experiments Using Web-Based Non-Parallel Corpora

The extraction experiments from the previous section are controlled experiments in which we only use limited amounts of parallel data for our extraction system. In this

section, we describe experiments in which the goal is to assess the applicability of our method to data that we mined from the Web.

We obtained comparable corpora from the Web by going to bilingual news websites (such as Al-Jazeera) and downloading news articles in each language independently. In order to get as many articles as possible, we used the web site's search engine to get lists of articles and their URLs, and then crawled those lists. We used the Agent-Builder tool (Ticrea and Minton 2003; Minton, Ticrea, and Beach 2003) for crawling. The tool can be programmed to automatically initiate searches with different parameters and to identify and extract the desired article URLs (as well as other information such as dates and titles) from the result pages. Table 5 shows the sources, time periods, and size of the datasets that we downloaded.

For the extraction experiments, we used dictionaries of high coverage, learned from all our available parallel training data. The sizes of these training corpora, measured in number of English tokens, are as follows:

- Arabic-English: 100M tokens out-of-domain data and 4.5M tokens in-domain data
- Chinese-English: 150M tokens out-of-domain data and 40M tokens in-domain data

We applied our extraction method on both the LDC-released Gigaword corpora and the Web-downloaded comparable corpora. For each language pair, we used the highest precision classifier from those presented in Section 3.4. In order to obtain data of higher quality, we didn't use all the sentences classified as parallel, but only those for which the probability computed by our classifier was higher than 0.70. Table 6 shows the amounts of extracted data, measured in number of English tokens. For Arabic-English, we were able to extract from the Gigaword corpora much more data than in our previous experiments (see Table 2), clearly due to the better dictionary. For Chinese-English, there was no increase in the size of extracted data (although the amount from Table 6 is smaller than that from Table 2, it counts only sentence pairs extracted with confidence higher than 0.70).

In the previous section, we measured, for our training corpora, their coverage of the test set (Tables 3 and 4). We repeated the measurements for the training data from Table 6 and obtained very similar results: using the additional extracted data improves coverage, especially for longer n-grams.

To give the reader an idea of the amount of data that is funneled through our system, we show in Figure 7 the sizes of the data processed by each of the system's

Table 5
Comparable corpora downloaded from the Web.

		Foreign		English	
Language pair	News agency and period	# articles	# tokens	# articles	# tokens
Arabic-English	People's Daily, 2001–2003 Al-Jazeera, 2003 Al-Hayat, 2003	70k	38M	50k	20M
Chinese-English	Voice of America, 2001–2003	25k	13M	36k	19M

Table 6
Size of the datasets extracted for the NIST 2004 MT evaluation.

Source	Arabic-English	Chinese-English
Gigaword	5.3M	7.2M
Web	1.4M	2.1M
Total	6.8M	9.3M

components during extraction from the Gigaword and Web-based Chinese-English comparable corpora. We use a dictionary learned from a parallel corpus on 190M English tokens and a classifier trained on instances generated from a parallel corpus of 220k English tokens. We start with a comparable corpus consisting of 500k Chinese articles and 600k English articles. The article selection step (Section 2.1) outputs 7.5M similar article pairs; from each article pair we generate all possible sentence pairs and obtain 2,400M pairs. Of these, less than 1% (17M) pass the candidate selection stage (Section 2.2) and are presented to the ME classifier. The system outputs 430k sentence pairs (9.5M English tokens) that have been classified as parallel (with probability greater than 0.7).

The figure also presents, in the lower part, the parameters that control the filtering at each stage.

- *best K results*: in the article selection stage (Section 2.1), for each foreign article we only consider the top K most similar English ones. In our experiments, K is set to 20.
- *date window*: when looking for possible article pairs, we only consider English articles whose publication dates fall within a window of 5 days around the publication date of the foreign one.
- *word overlap*: the word-overlap filter (Section 2.2) will discard sentence pairs that have less than a certain proportion of words in common (according to the bilingual dictionary). The value we use (expressed as a percentage of sentence length) is 50.
- *length ratio*: similarly, the word-overlap filter will discard pairs whose length ratio is greater than this value, which we set to 2.
- *decision threshold*: The ME classifier associates a probability with each of its predictions. Values above 0.5 indicate that the classifier considers the particular sentence pair to be parallel; the higher the value, the higher the classifier's confidence. Thus, in order to obtain higher precision, we can choose to define as parallel only those pairs for which the classifier probability is above a certain threshold. In the experiments from Section 4.1, we use the (default) threshold of 0.5, while in Section 4.2 we use 0.7.

5. Machine Translation Improvements

Our main goal is to extract, from an in-domain comparable corpus, parallel training data that improves the performance of an out-of-domain-trained SMT system. Thus,

we evaluate our extracted corpora by showing that adding them to the out-of-domain training data of a baseline MT system improves its performance.

5.1 Controlled Experiments

We first evaluate the extracted corpora presented in Section 4.1. The extraction system used to obtain each of those corpora made use of a certain *initial* out-of-domain parallel corpus. We train a *Baseline* MT system on that initial corpus. We then train another MT system (which we call *PlusExtracted*) on the initial corpus plus the extracted corpus. In order to compare the quality of our extracted data with that of human-translated data from the same domain, we also train an *UpperBound* MT system, using the initial corpus plus a corpus of in-domain, human-translated data. For each initial corpus, we use the same amount of human-translated data as there is extracted data (see Table 2). Thus, for each language pair and each *initial* parallel corpus, we compare 3 MT systems: Baseline, PlusExtracted, and UpperBound.

All our MT systems were trained using a variant of the alignment template model described in (Och 2003). Each system used two language models: a very large one, trained on 800 million English tokens, which is the same for all the systems; and a smaller one, trained only on the English side of the parallel training data for that particular system. This ensured that any differences in performance are caused only by differences in the training data.

The systems were tested on the news test corpus used for the NIST 2003 MT evaluation.[5] Translation performance was measured using the automatic BLEU evaluation metric (Papineni et al. 2002) on four reference translations.

Figures 8 and 9 show the BLEU scores obtained by our MT systems. The 95% confidence intervals of the scores computed by bootstrap resampling (Koehn 2004) are marked on the graphs; the delta value is around 1.2 for Arabic-English and 1 for Chinese-English.

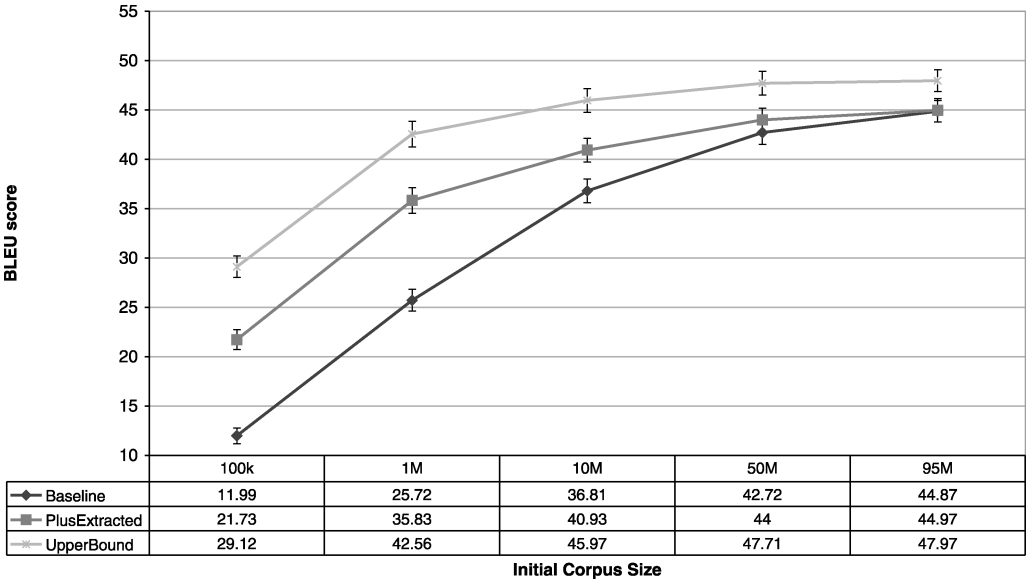
As the results show, the automatically extracted additional training data yields significant improvements in performance over most initial training corpora for both language pairs. At least for Chinese-English, the improvements are quite comparable to those produced by the human-translated data. And, as can be expected, the impact of the extracted data decreases as the size of the initial corpus increases.

In order to check that the classifier really does something important, we performed a few experiments without it. After the article selection step, we simply paired each foreign document with the best-matching English one, assumed they are parallel, sentence-aligned them with a generic sentence alignment method, and added the resulting data to the training corpus. The resulting BLEU scores were practically the same as the baseline; thus, our classifier does indeed help to discover higher-quality parallel data.

5.2 Non-Controlled Experiments

We also measured the MT performance impact of the extracted corpora described in Section 4.2. We trained a *Baseline* MT system on all our available (in-domain and

5 http://www.nist.gov/speech/tests/mt.



	100k	1M	10M	50M	95M
Baseline	11.99	25.72	36.81	42.72	44.87
PlusExtracted	21.73	35.83	40.93	44	44.97
UpperBound	29.12	42.56	45.97	47.71	47.97

Figure 8
MT performance improvements for Arabic-English.

out-of-domain) parallel data, and a *PlusExtracted* system on the parallel data plus the extracted in-domain data. Clearly, we have access to no UpperBound system in this case.

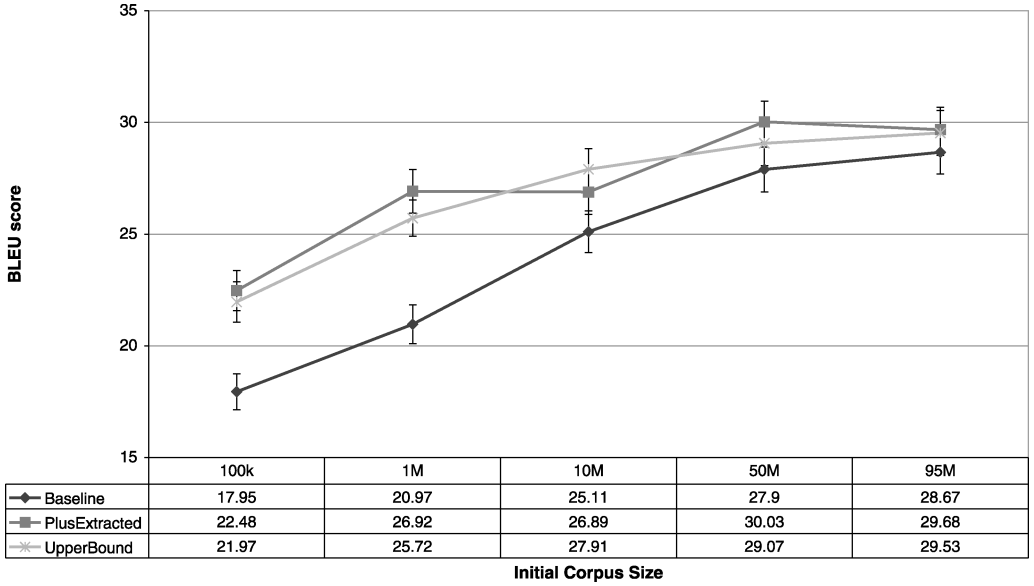
The results are presented in the first two rows of Table 7. Adding the extracted corpus lowers the score for the Arabic-English system and improves the score for the Chinese-English one; however, none of the differences are statistically significant. Since the baseline systems are trained on such large amounts of data (see Section 4.2), it is not surprising that our extracted corpora have no significant impact.

In an attempt to give a better indication of the value of these corpora, we used them alone as MT training data. The BLEU scores obtained by the systems we trained on them are presented in the third row of Table 7. For comparison purposes, the last line of the table shows the scores of systems trained on 10M English tokens of out-of-domain data. As can be seen, our automatically extracted corpora obtain better MT performance than out-of-domain parallel corpora of similar size. It's true that this is not a fair comparison, since the extracted corpora were obtained using all our available parallel data. The numbers do show, however, that the extracted data, although it was obtained automatically, is of good value for machine translation.

6. Bootstrapping

As can be seen from Table 2, the amount of data we can extract from our comparable corpora is adversely affected by poor dictionary coverage. Thus, if we start with very little parallel data, we do not make good use of the comparable corpora. One simple way to alleviate this problem is to bootstrap: after we've extracted some in-domain data, we can use it to learn a new dictionary and go back and extract again. Bootstrapping was also successfully applied to this problem by Fung and Cheung (2004).

We performed bootstrapping iterations starting from two very small corpora: 100k English tokens and 1M English tokens, respectively. After each iteration, we trained



	100k	1M	10M	50M	95M
Baseline	17.95	20.97	25.11	27.9	28.67
PlusExtracted	22.48	26.92	26.89	30.03	29.68
UpperBound	21.97	25.72	27.91	29.07	29.53

Figure 9
MT performance improvements for Chinese-English.

(and evaluated) an MT system on the initial data plus the data extracted in that iteration. We did not use any of the data extracted in previous iterations since it is mostly a subset of that extracted in the current iteration. We iterated until there were no further improvements in MT performance on our development data.

Figures 10 and 11 show the sizes of the data extracted at each iteration, for both initial corpus sizes. Iteration 0 is the one that uses the dictionary learned from the initial corpus. Starting with 100k words of parallel data, we eventually collect 20M words of in-domain Arabic-English data and 90M words of in-domain Chinese-English data.

Figures 12 and 13 show the BLEU scores of these MT systems. For comparison purposes, we also plotted on each graph the performance of our best MT system for that language pair, trained on all our available parallel data (Table 7).

As we can see, bootstrapping allows us to extract significantly larger amounts of data, which leads to significantly higher BLEU scores. Starting with as little as 100k English tokens of parallel data, we obtain MT systems that come within 7–10 BLEU points of systems trained on parallel corpora of more than 100M English tokens. This

Table 7
BLEU scores of the systems obtained using all available parallel data.

System	Arabic-English	Chinese-English
Baseline	49.22	33.77
Baseline plus extracted	48.54	34.38
Extracted only	41.2	28.04
Out-of-domain data	36.81	25.11

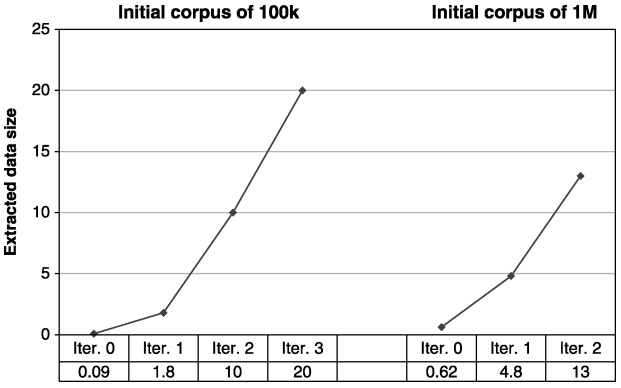


Figure 10
Sizes of the Arabic-English corpora extracted using bootstrapping, in millions of English tokens.

shows that using our method, a good-quality MT system can be built from very little parallel data and a large amount of comparable, non-parallel data.

7. Examples

We conclude the description of our method by presenting a few sentence pairs extracted by our system. We chose the examples by looking for cases when a given foreign sentence was judged parallel to several different English sentences. Figures 14 and 15 show the foreign sentence in Arabic and Chinese, respectively, followed by a human-produced translation in bold italic font, followed by the automatically extracted matching English sentences in normal font. The sentences are picked from the data sets presented in Section 4.2.

The examples reveal the two main types of errors that our system makes. The first type concerns cases when the system classifies as parallel sentence pairs that, although they share many content words, express slightly different meanings, as in Figure 15, example 7. The second concerns pairs in which the two sentences convey different amounts of information. In such pairs, one of the sentences contains a trans-

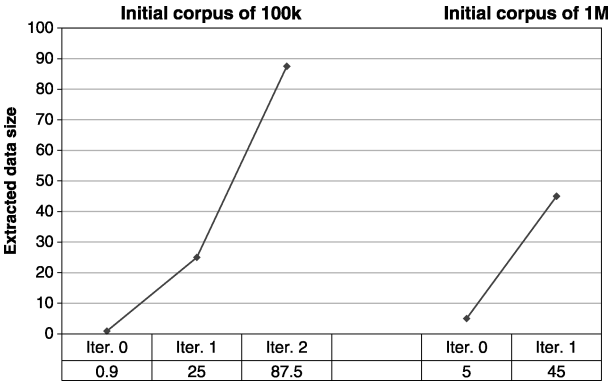


Figure 11
Sizes of the Chinese-English corpora extracted using bootstrapping, in millions of English tokens.

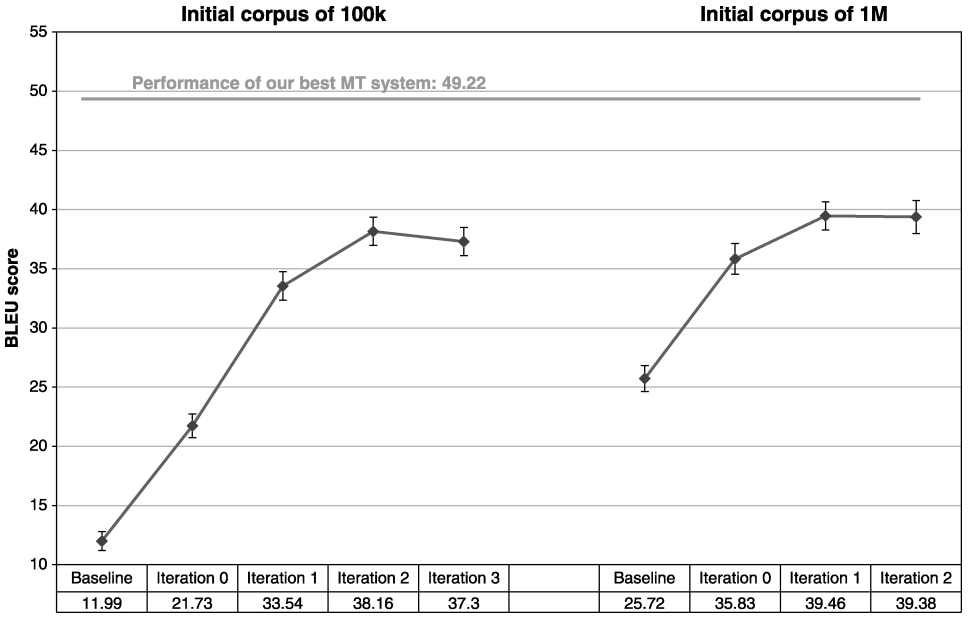


Figure 12
BLEU scores of the Arabic-English MT systems using bootstrapping.

lation of the other, plus additional (often quite long) phrases (Figure 15, examples 1 and 5).

These errors are caused by the noise present in the automatically learned dictionaries and by the use of a weak word alignment model for extracting the classifier

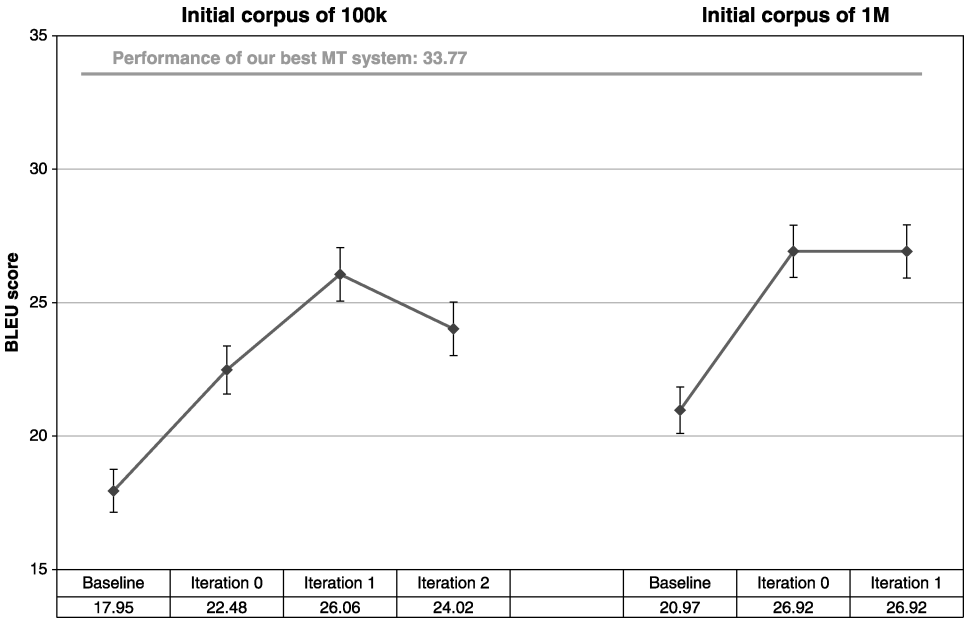


Figure 13
BLEU scores of the Chinese-English MT systems using bootstrapping.

1
وقال المتحدث باسم البنتاغون كينيث بايكون نحتاج إلى ضمانات تتيح لنا التأكد من أن في إمكاننا القيام بهذه المهمة باعتبارها مهمة إنسانية وليس مهمة قتالية
"We need guaranties that we will be able to carry out this mission given that it is a humanitarian mission and not a military mission," said Pentagon spokesman Kenneth Bacon.
- "We need assurances that we will be able to carry out this mission as a humanitarian mission not as a combat mission," said Pentagon spokesman Kenneth Bacon.
- "That final decision will depend on our ability to satisfy ourselves that we can carry out this mission as effectively as possible," said Pentagon spokesman Kenneth Bacon.

2
إلى أراضيها (انفصالي)في شمال العراق للحلول دون تسلل متمردي حزب العمال الكردستاني "منطقة أمنية"وكانت تركيا أعلنت منذ أسبوع عن نيتها إقامة
Meanwhile, Turkey has announced last week that it plans to set up a "security zone" in north of Iraq to stop the incursion of the Kurdish Workers Party rebels into its territory.
- Meanwhile, Turkey has announced plans to set up a border "security zone" inside northern Iraq to prevent incursions by Turkish Kurd PKK rebels based in the area.
- Ankara wants to create a temporary "security zone" along the border in northern Iraq to prevent infiltrations into southern Turkey by the outlawed Kurdistan Workers' Party (PKK).
- Baghdad accuses Tehran of backing the PKK, while Turkey has announced plans to set up a border "security zone" in northern Iraq to prevent infiltrations by Turkish Kurd rebels.

3
وصلت وزيرة الخارجية الأميركية مادلين أولبرايت إلى موسكو اليوم الخميس قادمة من لندن في زيارة تسعى خلالها إلى إقناع الرئيس بوريس يلتسين بأن ليس ثمة ما يدعو لقلق روسيا من مشروعات توسيع نطاق حلف شمال الأطلسي
US Secretary of State Madeleine Albright arrived to Moscow Thursday from London on a mission to convince President Boris Yeltsin that Russia should not worry about NATO's expansion plans.
- US Secretary of State Madeleine Albright arrived in Moscow Thursday on a mission to persuade President Boris Yeltsin that Russia has nothing to fear from NATO expansion.
- US Secretary of State Madeleine Albright attempted Friday to persuade President Boris Yeltsin that Russia has nothing to fear from NATO's eastward expansion, but there was no breakthrough.

4
للاجتماع مع نظيره البوليفي غونزالو سانشيز دي لوزادا (في شرق بوليفيا)ويزور الرئيس البيروفي البرتو فوجيموري بدءا من مساء اليوم الجمعة سانتا كروز
Peruvian President Alberto Fujimori is expected to visit Bolivian President Gonzalo Sanchez de Lozada in Santa Cruz (West of Bolivia) this Friday night.
- Peruvian President Alberto Fujimori held a working breakfast with Bolivian President Gonzalo Sanchez de Lozada in Santa Cruz, Bolivia, after a late-night meeting Friday.
- Another sign of a turning point in the hostage crisis was Fujimori's meeting with his Bolivian counterpart Gonzalo Sanchez de Lozada in Santa Cruz, Bolivia.

5
صراعات كثيرة ومعقدة على السلطة"وكان الناطق باسم البيت الأبيض مايكل ماكاري قال إن التطورات الأخيرة في العراق تدل على أنه يشهد
White House spokesman Michael McCurry said that recent events prove that Iraq is undergoing many complicated struggles for power.
- Responding to those reports, White House spokesman Michael McCurry said recent events showed a "great deal of complicated internal struggles for power. "
- White House spokesman Michael McCurry said recent events indicated a power struggle was underway in Iraq there was no sign Baghdad is preparing to attack Kuwait again.

6
المكسيك0 0 1 1
Mexico 0 0 1 1
- Zimbabwe 1 0 0 1
- Poland 1 0 0 1
- Holland 0 1 0 1
- Mexico 0 0 1 1
- Lithuania 0 0 1 1

Figure 14
Automatically extracted Arabic-English sentence pairs.

features. In an automatically learned dictionary, many words (especially the frequent, non-content ones) will have a lot of spurious translations. The IBM-1 alignment model takes no account of word order and allows a source word to be connected to arbitrarily many target words. Alignments computed using this model and a noisy, automatically learned, dictionary will contain many incorrect links. Thus, if two sentences share several content words, these incorrect links together with the correct links between the

1
(记者孙承斌)为期两天的 中英联合联络小组第 43 次会议今天上午在香港举行。
The Sino-British Joint Liaison Group(JLG) began its 43rd meeting today in Hong Kong and the meeting will last 2 days. (Reporter Chengbin Sun)
- The Sino-British Joint Liaison Group (JLG) began its 43rd meeting here Tuesday.
- The Sino-British Joint Liaison Group (JLG) ended its two-day meeting here Wednesday with a joint communique issued.

2
中罗两国 经贸合作具有很大的潜力和广阔的前景。
There is huge potential in economic and trade cooperation between Romania and China.
- The president said there's a huge potential in economic and trade cooperation between Romania and China.
- Roman said that Romania and China have many identical views on many international issues and the two countries have huge potentials in economic cooperation.

3
这是中国体操选手在本届奥运会上夺得的第一枚金牌 。
This is the first gold medal won by the Chinese gymnastic team at these Olympic games.
- This is China's first gold at the current Olympics.
- This is Ukraine's second gold medal in the Games.
- This is the first gold medal won by the Chinese team at the Atlanta Olympics.

4
俄罗斯总统叶利钦 13 日 表示,俄将向塔吉克斯坦政府提供技术和军事援助。
President Yeltsin said on the 13th that Russia will provide technical and military assistance to Tajikistan.
- President Boris Yeltsin said today that Russia will provide military technology and assistance to Tajikistan, where Tajik Islamic rebels have launched cross-border attacks from Afghanistan.
- Yeltsin stressed that it would be reasonable to render necessary technical and military assistance to Tajikistan.
- However, he added, it is also necessary to give Tajikistan technical and military aid.

5
中国残疾人艺术团成立于 1987 年,先后出访过近 30 个国家,被誉为美与友谊的使者。
The China Art Performance Troupe for the Disabled, which was founded in 1987, has visited almost 30 foreign countries and is called 'Messengers of Beauty and Friendship'.
- The troupe was founded in 1987 and has visited almost 30 foreign countries.
- The troupe, which was set up in 1987, has visited almost 30 foreign countries and is called by many overseas media as " Messengers of Beauty and Friendship".

6
里昂:
Lyon:
- Jamaica:
- At Lyon:
- Romania:
- Midfielders:
- Strikers:

7
叶利钦表示,俄罗斯将始终坚持一个中国的立场,不 会有任何偏离。
Yeltsin said that Russia will unswervingly stick to its one China policy with no deviations whatsoever.
- Yeltsin said that Russia will unswervingly stick to its one China policy with no deviations whatsoever.
- China expressed its heartfelt gratitude to Russia for the principle of "one China" it has consistently upheld.

Figure 15
Automatically extracted Chinese-English sentence pairs.

common content words will yield an alignment good enough to make the classifier judge the sentence pair as parallel.

The effect of the noise in the dictionary is even more clear for sentence pairs with few words, such as Figure 14, example 6. The sentences in that example are tables of soccer team statistics. They are judged parallel because corresponding digits align

to each other, and according to our dictionary, the Arabic word for "Mexico" can be translated as any of the country names listed in the example.

These examples also show that the problem of finding only true translation pairs is hard. Two sentences may share many content words and yet express different meanings (see Figure 14, example 1). However, our task of getting useful MT training data does not require a perfect solution; as we have seen, even such noisy training pairs can help improve a translation system's performance.

8. Related Work

While there is a large body of work on bilingual comparable corpora, most of it is focused on learning word translations (Fung and Yee 1998; Rapp 1999; Diab and Finch 2000; Koehn and Knight 2000; Gaussier et al. 2004). We are aware of only three previous efforts aimed at discovering parallel sentences. Zhao and Vogel (2002) describe a generative model for discovering parallel sentences in the Xinhua News Chinese-English corpus. Utiyama et. al (2003) use cross-language information retrieval techniques and dynamic programming to extract sentences from an English-Japanese comparable corpus. Fung and Cheung (2004) present an extraction method similar to ours but focus on "very-non-parallel corpora," aggregations of Chinese and English news stories from different sources and time periods.

The first two systems extend algorithms designed to perform sentence alignment of parallel texts. They start by attempting to identify similar article pairs from the two corpora. Then they treat each of those pairs as parallel texts and align their sentences by defining a sentence pair similarity score and use dynamic programming to find the least-cost alignment over the whole document pair.

In the article pair selection stage, the researchers try to identify, for an article in one language, the best matching article in the other language. Zhao and Vogel (2002) measure article similarity by defining a generative model in which an English story generates a Chinese story with a given probability. Utiyama et al. (2003) use the BM25 (Robertson and Walker 1994) similarity measure.

The two works also differ in the way they define the sentence similarity score. Zhao and Vogel (2002) combine a sentence length model with an IBM Model 1-type translation model. Utiyama et al. (2003) define a score based on word overlap (i.e., number of word pairs from the two sentences that are translations of each other), which also includes the similarity score of the article pair from which the sentence pair originates.

The performance of these approaches depends heavily on the ability to reliably find similar document pairs. Moreover, comparable article pairs, even those similar in content, may exhibit great differences at the sentence level (reorderings, additions, etc). Therefore, they pose hard problems for the dynamic programming alignment approach.

In contrast, our method is more robust. The document pair selection part plays a minor role; it only acts as a filter. We do not attempt to find the best-matching English document for each foreign one, but rather a set of similar documents. And, most importantly, we are able to reliably judge each sentence pair in isolation, without need for context. On the other hand, the dynamic programming approach enables discovery of many-to-one sentence alignments, whereas our method is limited to finding one-to-one alignments.

The approach of Fung and Cheung (2004) is a simpler version of ours. They match each foreign document with a set of English documents, using a threshold on their

cosine similarity. Then, from each document pair, they generate all possible sentence pairs, compute their cosine similarity, and apply another threshold in order to select the ones that are parallel. Using the set of extracted sentences, they learn a new dictionary, try to extend their set of matching document pairs (by looking for other documents that contain these sentences), and iterate.

The evaluation methodologies of these previous approaches are less direct than ours. Utiyama et al. (2003) evaluate their sentence pairs manually; they estimate that about 90% of the sentence pairs in their final corpus are parallel. Fung and Cheung (2004) also perform a manual evaluation of the extracted sentences and estimate their precision to be 65.7% after bootstrapping. In addition, they also estimate the quality of a lexicon automatically learned from those sentences. Zhao and Vogel (2002) go one step further and show that the sentences extracted with their method improve the accuracy of automatically computed word alignments, to an F-score of 52.56% over a baseline of 46.46%. In a subsequent publication, Vogel (2003) evaluates these sentences in the context of an MT system and shows that they bring improvement under special circumstances (i.e., a language model constructed from reference translations) designed to reduce the noise introduced by the automatically extracted corpus. We go even further and demonstrate that our method can extract data that improves end-to-end MT performance without any special processing. Moreover, we show that our approach works even when only a limited amount of initial parallel data (i.e., a low-coverage dictionary) is available.

The problem of aligning sentences in comparable corpora was also addressed for monolingual texts. Barzilay and Elhadad (2003) present a method of aligning sentences in two comparable English corpora for the purpose of building a training set of text-to-text rewriting examples. Monolingual parallel sentence detection presents a particular challenge: there are many sentence pairs that have low lexical overlap but are nevertheless parallel. Therefore pairs cannot be judged in isolation, and context becomes an important factor. Barzilay and Elhadad (2003) make use of contextual information by detecting the topical structure of the articles in the two corpora and aligning them at paragraph level based on the topic assigned to each paragraph. Afterwards, they proceed and align sentences within paragraph pairs using dynamic programming. Their results show that both the induced topical structure and the paragraph alignment improve the precision of their extraction method.

A line of research that is both complementary and related to ours is that of Resnik and Smith (2003). Their STRAND Web-mining system has a purpose that is similar to ours: to identify translational pairs. However, STRAND focuses on extracting pairs of parallel Web pages rather than sentences. Resnik and Smith (2003) show that their approach is able to find large numbers of similar document pairs. Their system is potentially a good way of acquiring comparable corpora from the Web that could then be mined for parallel sentences using our method.

9. Discussion

The most important feature of our parallel sentence selection approach is its robustness. Comparable corpora are inherently noisy environments, where even similar content may be expressed in very different ways. Moreover, out-of-domain corpora introduce additional difficulties related to limited dictionary coverage. Therefore, the ability to reliably judge sentence pairs in isolation is crucial.

Comparable corpora of interest are usually of large size; thus, processing them requires efficient algorithms. The computational processes involved in our system are

quite modest. All the operations necessary for the classification of a sentence pair (filter, word alignment computation, and feature extraction) can be implemented efficiently and scaled up to very large amounts of data. The task can be easily parallelized for increased speed. For example, extracting data from 600k English documents and 500k Chinese documents (Section 4.2) required only about 7 days of processing time on 10 processors.

The data that we extract is useful. Its impact on MT performance is comparable to that of human-translated data of similar size and domain. Thus, although we have focused our experiments on the particular scenario where there is little in-domain training data available, we believe that our method can be useful for increasing the amount of training data, regardless of the domain of interest.

As we have shown, this could be particularly effective for language pairs for which only very small amounts of parallel data are available. By acquiring a large comparable corpus and performing a few bootstrapping iterations, we can obtain a training corpus that yields a competitive MT system.

We suspect our approach can be used on comparable corpora coming from any domain. The only domain-dependent element of the system is the *date window* parameter of the article selection stage (Figure 7); for other domains, this can be replaced with a more appropriate indication of where the parallel sentences are likely to be found. For example, if the domain were that of technical manuals, one would cluster printer manuals and aircraft manuals separately. It is important to note that our work assumes that the comparable corpus does contain parallel sentences (which is the case for our data). Whether this is true for comparable corpora from other domains is an empirical question outside the scope of this article; however, both our results and those of Resnik and Smith (2003) strongly indicate that good data is available on the Web.

Lack of parallel corpora is a major bottleneck in the development of SMT systems for most language pairs. The method presented in this paper is a step towards the important goal of automatic acquisition of such corpora. Comparable texts are available on the Web in large quantities for many language pairs and domains. In this article, we have shown how they can be efficiently mined for parallel sentences.

Acknowledgments

This work was supported by DARPA-ITO grant NN66001-00-1-9814 and NSF grant IIS-0326276. The experiments were run on University of Southern California's high-performance computer cluster HPC (http://www.usc.edu/hpcc). We would like to thank Hal Daumé III, Alexander Fraser, Radu Soricut, as well as the anonymous reviewers, for their helpful comments. Any remaining errors are of course our own.

References

Barzilay, Regina and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pages 25–32, Sapporo, Japan.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

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

Darroch, J. N. and D. Ratcliff. 1974. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:95–144.

Davis, Mark W. and Ted E. Dunning. 1995. A TREC evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference*, pages 483–498, Gaithersburg, MD.

Diab, Mona and Steve Finch. 2000. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access*, Paris, France.

Diab, Mona and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 255–262, Philadelphia.

Echihabi, Abdessamad and Daniel Marcu. 2003. A noisy-channel approach to question answering. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 16–23, Sapporo, Japan.

Fung, Pascale and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and EM. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, pages 57–63, Barcelona, Spain.

Fung, Pascale and Kenneth Ward Church. 1994. Kvec: A new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, pages 1096–1102, Kyoto.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 414–420, Montreal.

Gale, William A. and Kenneth Ward Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Berkeley, CA.

Gaussier, Eric, Jean-Michel Renders, Irina Matveeva, Cyril Goutte, and Herve Dejean. 2004. A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 527–534, Barcelona, Spain.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain.

Koehn, Philipp and Kevin Knight. 2000. Estimating word translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 711–715, Austin, TX.

Malouf, Robert. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Sixth Conference on Natural Language Learning*, Taipei, Taiwan.

Melamed, Dan I. 1997. A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Madrid, Spain.

Melamed, Dan I. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130.

Minton, Steven N., Sorinel I. Ticrea, and Jennifer Beach. 2003. Trainability: Developing a responsive learning system. In *IJCAI Workshop on Information Integration on the Web*, pages 27–32, Acapulco, Mexico.

Moore, Robert C. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas*, pages 135–144, Tiburon, CA.

Moore, Robert C. 2004. Improving IBM word-alignment model 1. In *42nd Annual Meeting of the Association for Computational Linguistics*, pages 519–526, Barcelona, Spain.

Munteanu, Dragos Stefan, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association For Computational Linguistics*, pages 265–272, Boston, MA.

Oard, Douglas W. 1997. Cross-language text retrieval research in the USA. In *Third DELOS Workshop on Cross-Language Information Retrieval*, pages 1–10, Zurich, Switzerland.

Och, Franz Josef. 2003. Minimum error rate training for statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, Franz Josef and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia.

Och, Franz Joseph and Hermann Ney. 2003. A systematic comparison of various

statistical alignment models. *Computational Linguistics*, 29(1):19–51.
Ogilvie, Paul and Jamie Callan. 2001. Experiments using the Lemur toolkit. In *Proceedings of the Tenth Text Retrieval Conference*, pages 103–108, Gaithersburg, MD.
Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia.
Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 519–526, College Park, MD.
Resnik, Philip and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380, September.
Robertson, E. and S. Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual ACM SIGIR*, pages 232–241, Dublin, Ireland.
Ticrea, Sorinel I. and Steven Minton. 2003. Inducing web agents: Sample page management. In *Proceedings of the International Conference on Information and Knowledge Engineering*, pages 399–403, Las Vegas, NV, June.
Utiyama, Masao and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan.
Vogel, Stephan. 2003. Using noisy bilingual data for statistical machine translation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 175–178, Budapest, Hungary.
Wu, Dekai. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 80–87, Las Cruces, NM.
Yarowsky, David and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the 2nd Meeting of the North American Association for Computational Linguistics*, pages 200–207, Pittsburgh, PA.
Yarowsky, David, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, pages 161–168, San Diego, CA.
Zhao, Bing and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *2002 IEEE International Conference on Data Mining*, pages 745–748, Maebashi City, Japan.