Improved Machine Translation Performance via Parallel Sentence Extraction from Comparable Corpora

Dragos Stefan Munteanu, Alexander Fraser, Daniel Marcu
University of Southern California
Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA, 90292
{dragos,fraser,marcu}@isi.edu

Abstract

We present a novel method for discovering parallel sentences in comparable 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, Gigaword, Arabic and English newspaper corpora. We evaluate the quality of the extracted data by showing it improves the performance of a baseline statistical machine translation system.

1 Introduction

Parallel texts are an important resource in many NLP applications. Unfortunately, they are often scarce resources: limited in size, language coverage and language register. Much more readily available are comparable corpora that, while not parallel in the strict sense, are closely related and convey the same information. The best example of such texts are the multilingual news feeds produced by several news agencies (Agence France Presse, Xinhua News, etc).

In the field of statistical machine translation (SMT), it is often the case that the available parallel training data comes primarily from one domain (e.g. parliamentary proceedings), but the translation system one builds needs to perform well on a different domain (e.g. news). This is hard, because the system will perform poorly in a domain that is different from the one it was trained on.

There are several ways to deal with this problem. The solution that we investigate is to extract parallel sentences from comparable corpora from the domain of interest, and use the extracted sentences as additional training material. In our experiments, the training genre is United Nations English-Arabic data. We refer to this as *in-domain* data. The test genre is that of Arabic news stories translated into English; we call this *out-of-domain* data.

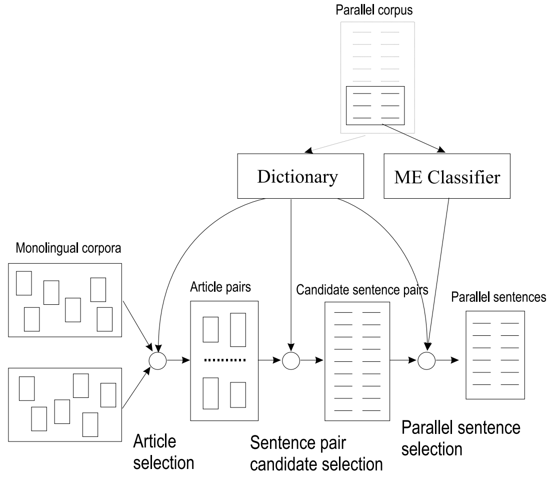

Figure 1: Parallel Sentence Extraction System

Several other researchers have used comparable corpora to extract bilingual information (mostly word translations) with the implicit goal of building better translation systems. However, to our knowledge, none has shown empirically that this is possible. In this paper, we show that sentences extracted with our method improve the end-to-end performance of a statistical machine translation (MT) system on out-of-domain test data.

Figure 1 illustrates our approach to the problem of parallel sentence extraction. Starting with two large monolingual corpora (a comparable corpus) divided into articles, we begin by selecting pairs of similar articles (Section 2.1). From each such pair, we take 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 2.4) that decides whether the sentences in each pair are mutual translations of each other.

The resources required by the system are minimal: a dictionary and a small amount of parallel data which is used for training the ME classifier. The dictionary used in our experiments was obtained automatically from an in-domain parallel corpus[1]; thus, the input to our system consisted only of small amounts of parallel data.

In the next section we describe in more detail each component of the system. We then present our experiments and results, talk about related work, and conclude.

2 Extracting Parallel Sentences from Comparable Corpora

2.1 Article Selection

Our comparable corpus consists of two large monolingual news corpora, one written in English and one in Arabic (described in more detail in Section 3.2). The parallel sentence extraction process begins by selecting, for each Arabic article, English articles that are likely to contain sentences that are parallel to those in the Arabic document.

Our approach emphasizes recall rather than precision. For each Arabic document, we do not attempt to find the best matching English document, but rather a set of English documents that are similar to the Arabic one. 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 IR engine InQuery (Callan et al., 1995). We index all the English documents into a database, and create a query for each Arabic document. We take the top 5 translations (according to our probabilistic dictionary) of each word in the document, and create a query using InQuery's *wsum* (weighted sum) operator (using as weights the translation probabilities). We run the query and retrieve the top 100 English documents.

We consider it likely that documents with similar content have publication dates that are close to each other. Thus, from the top 100 English documents returned by InQuery, we actually keep only those published within a window of 5 days around the publication date of the Arabic query document.

2.2 Candidate Sentence Pair Selection

From each Arabic 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 2. It then checks that at least half the words in each sentence have a translation in the other sentence. Pairs that do not fulfill these two conditions are discarded. The others are passed on to the parallel sentence selection stage.

This step removes much of the noise (i.e. pairs of non-parallel sentences) introduced by our recall-oriented document selection procedure. It also removes good pairs, which 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.

2.3 Parallel sentence selection

For each candidate sentence pair, we need to reliably decide whether the two sentences in the pair are mutual translations. This is achieved by a Maximum Entropy (ME) classifier, which is the core component of our system. Those pairs that are classified as being translations of each other constitute the output of our system.

We first explain the intuition behind the design of the classifier, and then present the implementation details.

2.4 A Maximum Entropy Classifier For Parallel Sentence Detection

In the 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 number 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*.

The ME principle suggests that 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:

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

where c is the class ("parallel" or "not parallel"), $Z(sp)$ is a normalization factor, and f_i are the feature functions. 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 algorithms such as GIS (Darroch and Ratcliff, 1974) or its improved version IIS (Berger et al., 1996).

For our particular 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

[1] If such a resource is unavailable, other dictionaries can be used

two sentences in different languages specifies which words in one sentence are translations of 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 2 and 3 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 was produced by a human, while the one on the right was computed automatically. As can be seen from the gloss next to the Arabic words, the sentences in Figure 2 are parallel while the sentences in Figure 3 are not.

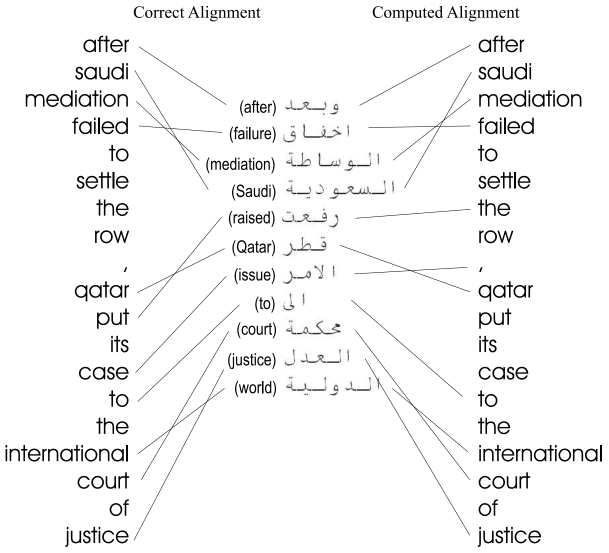

Figure 2: Alignments between two parallel sentences

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 (Figure 3), due to noisy dictionary entries and to 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 (i.e. Arabic word *at* in Figure 3) is indicative of non-parallelism. Most

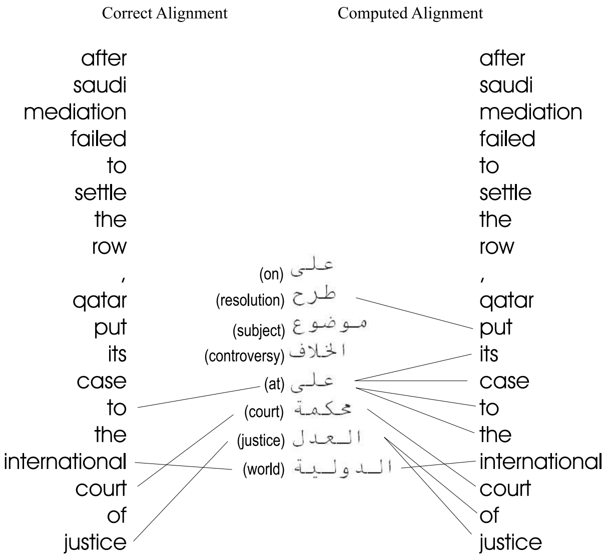

Figure 3: Alignments between two non-parallel sentences

likely, these connections were produced because of a lack of better alternatives.

Another feature of interest is the presence of long *contiguous 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. 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 2: the English strings *after saudi mediation failed* or *to the international court of justice* together with their Arabic counterparts. Long contiguous spans are indicative of parallelism, since they suggest that the two sentences have long phrases in common.

If the dictionary is probabilistic, we can define the *alignment score* to be the (normalized) product of the translation probabilities of the connected word pairs. This score is also an indicative factor: a pair of non-parallel sentences should have connections of lower probability.

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;

Alignment features:

- percentage and number of words that have no connection;

- the top 3 largest fertilities;
- length of the longest contiguous span;
- alignment score;

2.5 Implementation

In order to compute word alignments, we use the IBM Model 1 (Brown et al., 1993). We chose this model because it is simple, efficient, and has shown in our experiments to have good discriminative power.

For a source sentence f and a target sentence e, the joint likelihood of sentence f and an alignment a given sentence e is defined as:

$$P(f,a|e) = \frac{e}{(l+1)^m} \prod_{j=1}^{m} t(f_j|e_{a_j})$$

where m is the length of the source sentence, l is the length of the target sentence, and $e = P(m|e)$.

This is basically the normalized product of the translation probabilities of all the links in the alignment. We compute the best alignment according to this model, which is the one that maximizes the product term:

$$\hat{a} = argmax_a(P(f,a|e)) = argmax_a(\prod_{j=1}^{m} t(f_j|e_{a_j}))$$

Following (Och and Ney, 2003), we compute one alignment for each translation direction ($f \rightarrow e$ and $e \rightarrow f$) and then combine them together. Och and Ney present 3 combination methods: *intersection*, *union*, and *refined* which is a form of intersection expanded with certain additional neighboring links.

Thus, for each sentence pair we compute 5 alignments (2 IBM-Model-1 plus 3 combinations), and extract one set of general features and 5 sets of alignment features (as described in the previous section).

We train the parameters of the model on instances obtained from a small parallel corpus. We take all possible bilingual sentence pairs from the corpus, put them through the word overlap filter described in Section 2.2, and create training instances from those that pass the filter. We compute the values of the classifier parameters using the YASMET[2] implementation of the GIS algorithm.

3 Results

We assess the performance of our parallel sentence extractor in the context of an end-to-end Arabic-English MT system. We used two parallel corpora for the MT system: an in-domain corpus of 2.7 million sentences of United Nations Arabic-English data (63M English words), and an out-of-domain corpus of 17,000 sentences (420k English words) of Arabic news data translated into English.

We begin by presenting performance results for the ME classifier. We then show the results of applying our sentence extraction method on two large (gigaword) monolingual collections of news articles, under two different experimental conditions.

The first experiment is designed to assess the quality of the data extracted with our system. Using a dictionary trained on all our in-domain data, and two classifiers (trained on different datasets), we automatically extract two parallel corpora. We perform a thorough evaluation of these corpora, showing they yield improvements in MT performance over baselines of various sizes.

The second experiment addresses the case when only a small amount of parallel data is available. We use a restricted amount of parallel data for building the extraction system, and show the MT performance gains obtained by adding the extracted corpus to that initial data.

3.1 Evaluation of the ME Classifier

As described in Section 2.3, all sentence pairs seen by the classifier are first passed through a word overlap filter (Section 2.2). Thus, training (or testing) a classifier requires a dictionary for the filter and a parallel corpus for generating training (or test) instances.

We want to examine the impact that both the dictionary coverage and the training corpus domain have upon the performance of the classifier. We therefore prepared:

- 5 dictionaries of various sizes (i.e. trained on in-domain parallel corpora of various sizes), and
- two parallel training corpora, one in-domain and the other out-of-domain, 5000 sentence pairs each

and trained classifiers using all combinations of dictionary and training corpora.

In order to train a classifier, we take the 5000-sentence training corpus, generate all possible sentence pairs, and pass them through the word overlap filter. From the pairs that pass the filter, for each English sentence we keep as ME training instances its Arabic equivalent (if it passes the filter) and at most one Arabic non-parallel. This ensures we have a balanced training set. Depending on the domain of the ME training corpus and the size of the filter's dictionary, we obtain between 6000 and 9500 training instances, out of which between 40% and 60% are positive.

We test our classifiers by generating test instances from 7000 held-out, out-of-domain parallel sentence pairs. We again generate all possible sentence

[2] http://www.isi.edu/~och/YASMET.html

pairs and pass them through the word overlap filter. Those that are accepted constitute the test instances. For each classifier, we use the same dictionary for both training and testing. We obtain between 5000 and 7000 instances, out of which around 70% are positive.

Table 1 shows the precision, recall, and F-score for the two sets of classifiers; note that these numbers are computed with respect to those pairs that have passed the word overlap filter. The row indices represent the amount of data used to learn the probabilistic dictionary (measured in number of English words), and the columns correspond to the two training corpora.

	in-domain training			out-of-domain training		
	P	R	F	P	R	F
5M	97	49	65.11	92	90	91.00
10M	97	48	64.22	93	90	91.50
25M	97	45	61.48	94	88	90.90
50M	97	50	65.99	94	88	90.90
63M	98	46	62.61	94	89	91.43

Table 1: Classifier performance on out-of-domain test data

These results show that the classifier (in conjunction with the filter) is robust with respect to dictionary coverage. They also show that the classifier's performance improves significantly if its parameters are trained on data from the same domain we apply it to.

One interesting point is that the precision of the in-domain-trained classifier is constantly higher than that of the out-of-domain classifier. One possible explanation for this is that our in-domain data, the UN corpus, has very high-quality, literal translations; thus, the classifier trained on this data learns to be more selective.

Evaluations performed using different subsets of features show that 99% 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 for real data, differences between parallel and non-parallel pairs are less clear than for our test data, and can no longer be accounted for only by counting the linked words; thus, the other features should become more important.

3.2 The Gigaword Corpora

The comparable corpus used in our experiments is built from two large monolingual news corpora: the English Gigaword Corpus[3] and the Arabic Gigaword Corpus[4], released by the Linguistic Data Consortium. Each corpus contains news stories produced over several years by various news agencies.

We only consider the common parts of the two corpora: stories in both corpora that come from the same news agency and the same time period. Thus, our comparable corpus consists of:

- stories from Agence France Press, published in 1994-1997 and the first half of 2002
- stories from The Xinhua News Agency, published in the second half of 2001

In total, the English side has 670k articles and 190M tokens, while the Arabic side has 280k articles and 50M tokens.

The texts were pre-processed in a manner consistent with the processing of the training data of our MT system. The English was sentence splitted using MXTerminator[5], tokenized and lowercased. The Arabic was converted to the CP1256 encoding (it was originally in UTF8), normalized, tokenized, and sentence splitted (using a simple heuristic that splits at every dot).

3.3 Evaluation of the Sentences Extracted Using All Resources

This section presents the evaluation of the best data we could extract from the Gigaword corpora presented above, using either only in-domain parallel data or both in-domain and out-of-domain parallel data for training the ME classifier.

As shown in Figure 1, our sentence extraction method requires a dictionary and a classifier. We extracted two sets of sentence pairs, using the same dictionary (trained on all the available in-domain parallel data) and two different classifiers: the best in-domain-trained classifier and the best out-of-domain-trained classifier, according to the results presented in Table 1.

Using the in-domain classifier we extracted from our comparable corpus 63,000 sentence pairs (approximately 1.7M English tokens and 1.61M Arabic tokens), which we will refer to as the *small corpus*. The out-of-domain classifier produced 200,000 pairs (6M English tokens and 5.3M Arabic tokens), which we will refer to as the *large corpus*. The difference between the sizes of the extracted corpora is consistent with the classifier evaluations presented in Section 3.1: the classifier trained on in-domain data has significantly lower recall.

We evaluate the quality of these corpora by adding them to training data sets of various sizes. Thus, we prepared in-domain parallel corpora of different

[3] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T05`

[4] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T12`

[5] `http://www.cis.upenn.edu/~adwait/statnlp.html`

	In-domain	Out-of-domain		
		Extracted		high-quality
		small	large	
0		95.05/64.76/35.76/21.79	97.63/72.88/44.86/29.08	91.70/45.33/15.71/5.92
5M	91.23/44.58/11.66/2.52	97.39/69.33/38.04/22.37	98.19/75.42/46.16/29.49	96.09/58.10/21.49/7.28
10M	92.87/49.89/14.47/3.24	97.70/70.86/39.05/22.68	98.37/76.30/46.80/29.68	96.76/61.41/23.45/7.87
25M	94.23/56.03/18.98/4.61	98.23/73.12/40.73/23.30	98.64/77.75/48.01/30.18	97.52/65.63/26.83/8.96
40M	95.08/59.64/21.97/5.63	98.53/74.60/42.09/23.73	98.86/78.70/48.95/30.51	98.07/68.06/29.21/9.89
63M	95.45/63.06/24.66/6.89	98.76/76.12/43.52/24.43	99.00/79.81/50.00/31.04	98.35/70.50/31.36/10.97

Table 2: Training corpus coverage: percentage of unigrams, bigrams, trigrams and 4-grams from the test corpus that are also present in the training corpus

sizes, and for each of them we trained and evaluated 4 SMT systems:

- the *baseline* system trained only on the in-domain data;
- the *small* system trained on the in-domain data plus the small extracted corpus;
- the *large* system trained on the in-domain data plus the large extracted corpus;
- the *high-quality* system trained on the in-domain data plus 10,000 sentences (250k tokens) of high-quality out-of-domain parallel data [6]).

An important indicator of the usefulness of MT training data is its "coverage" of the test data, i.e. the percentage of n-grams from the test corpus that are also in the training corpus. We present in Table 2 the coverage of each of the training corpora used in these experiments. Each cell contains 4 numbers, which 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 corpora from the same domain as the test data have much greater coverage than those from a different domain, regardless of the sizes.

We trained the translation systems using a variant of the model described in (Och, 2003). We tested them on the out-of-domain test corpus used for the TIDES 2002 MT evaluation. The translation performance was measured using the automatic BLEU (Papineni et al., 2002) evaluation metric, on 4 reference translations.

Table 3 shows the BLEU scores of our systems. The row indices are the sizes of the baseline parallel corpora measured in million English words, and the columns correspond to the 4 systems that were trained for each baseline size. The scores printed in boldface are different from those to their left at a statistically significant level.

	In-domain	Out-of-domain		
		Extracted		high-quality
		small	large	
0		34.81	35.17	35.42
5M	32.58	**38.06**	**36.91**	**38.79**
10M	33.98	**37.44**	37.66	**39.66**
25M	37.85	**39.92**	40.02	**41.88**
40M	38.99	**40.85**	40.27	**41.60**
63M	39.43	**40.61**	41.01	**42.63**

Table 3: BLEU scores

The BLEU scores from Table 3 are again consistent with the classifier precision results from section 3.1, and show that the small corpus is of higher quality: it yields as much improvement as the large corpus, although it is less than half its size. This shows that our method can extract good quality out-of-domain parallel sentences using only in-domain resources.

Still, neither of our extracted corpora is as helpful as the high-quality out-of-domain data, despite being significantly larger, and despite having better vocabulary coverage. This is most likely due to the fact that translations automatically extracted from comparable corpora are inherently noisy (there are reformulations, small differences in content, as well as incorrectly paired sentences), and current statistical translation models have only limited abilities to deal with noisy translations.

3.4 Evaluation of Sentences Extracted Using Limited Resources

In this section we evaluate the applicability of our approach in situations where only a small amount of parallel data is available. Thus, given a parallel corpus of a certain size, we use only that corpus for training the dictionary and the classifier needed by our system. We then apply the system to the two Gigaword comparable corpora.

We prepared baseline parallel corpora of various sizes, and used each of them to extract parallel sentences as described above. Table 4 presents the sizes of the extracted corpora, in number of sentences and

[6] Part of the Arabic-English parallel news data available for the 2003 NIST MT Evaluation, `http://www.nist.gov/speech/tests/mt/resources/index.htm`

tokens. The leftmost column indicates the size of the initial (baseline) parallel corpus.

	Sent. pairs	English tokens	Arabic tokens
5M	46k	1.34M	1.24M
10M	50k	1.40M	1.30M
25M	60k	1.70M	1.59M
40M	61k	1.71M	1.60M
63M	63k	1.73M	1.61M

Table 4: Sizes of extracted corpora

Table 5 shows the BLEU scores obtained with our extracted corpora. The *baseline* scores are those of the systems trained on the initial corpus. The *extracted* scores are obtained by adding to the MT training data the sentences extracted using only the baseline parallel corpus. The *high-quality* scores are the same as in Table 3. The scores printed in boldface are different from those to their left at a statistically significant level.

	In-domain	Out-of-domain	
	baseline	extracted	high-quality
5M	32.58	**37.62**	**38.79**
10M	33.98	**38.33**	**39.66**
25M	37.85	**39.72**	**41.88**
40M	38.99	**40.38**	**41.60**
63M	39.43	**40.61**	**42.63**

Table 5: BLEU scores, limited data

4 Related Work

While there is a large body of work on bilingual comparable corpora, most of it is focused on extracting word translations (Rapp, 1999; Diab and Finch, 2000; Fung and Yee, 1998; Koehn and Knight, 2000). We are aware of only two previous efforts to discover parallel sentences. Zhao et. al (2002) describe a generative model for discovering parallel sentences in the Xinhua Chinese-English corpus. Utiyama et. al (2003) use CLIR techniques and dynamic programming to extract sentences from an English-Japanese comparable corpus.

Both systems extend algorithms designed to perform sentence alignment of parallel texts. They define a sentence alignment score, and use dynamic programming to find the best sentence alignment between a pair of documents that are hypothesized to be similar. Thus, performance depends heavily on the ability to 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. 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 evaluation methodologies used by Utiyama et. al (2003) and Zhao et. al (2002) are less direct than ours. Utiyama et. al. perform a manual evaluation; the complete lack of English-Japanese parallel corpora leaves them no alternative. Zhao et. al go one step further, and show that the sentences extracted with their method improve the accuracy of automatically computed word alignments. 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. language model constructed from reference translations), designed to reduce the noise introduced by the automatically extracted corpus. We go further and demonstrate that our method can extract data which improves MT performance over baselines of various sizes, without any special processing. Moreover, we show that our approach works even when a limited amount of parallel data is available.

The problem of aligning sentences in comparable corpora was also addressed for monolingual texts. Barzilay et. al (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, context becomes an crucial factor.

5 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 sizes; 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, feature extraction) can be implemented efficiently, and can scale up to large amounts of data.

The data that we extract is useful. Even without using out-of-domain data, our system is able to produce a corpus that significantly improves out-of-domain translation performance. The extracted sentences also contain words unknown to our dictionary; by training an MT system on these automatically extracted sentences we learn new dictionary entries.

Lack of parallel corpora is a major bottleneck in MT research. 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 paper, we have shown how they can be efficiently and robustly mined for parallel sentences.

Acknowledgements

This work was supported by DARPA grant N66001-00-1-9814.

References

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

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Peter F. Brown, John Cocke, Stephen 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.

Peter F. Brown, 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, June.

J.P. Callan, W.B. Croft, and J. Broglio. 1995. TREC and Tipster experiments with InQuery. *Information Processing and Management*, 31(3):327–343.

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

Mona Diab 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.

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

Abdessamad Echihabi 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.

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

Philipp Koehn 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, USA.

Franz Joseph Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1).

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics*, 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 the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the Conference of the Association for Computational Linguistics*, pages 519–526, College Park, MD, USA.

Masao Utiyama 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.

Stephan Vogel. 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.

Bing Zhao 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.