# Extracting Parallel Phrases from Comparable Data

**Sanjika Hewavitharana** and **Stephan Vogel**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{sanjika,vogel+}@cs.cmu.edu

## Abstract

Mining parallel data from comparable corpora is a promising approach for overcoming the data sparseness in statistical machine translation and other NLP applications. Even if two comparable documents have few or no parallel sentence pairs, there is still potential for parallelism in the sub-sentential level. The ability to detect these phrases creates a valuable resource, especially for low-resource languages. In this paper we explore three phrase alignment approaches to detect parallel phrase pairs embedded in comparable sentences: the standard phrase extraction algorithm, which relies on the Viterbi path; a phrase extraction approach that does not rely on the Viterbi path, but uses only lexical features; and a binary classifier that detects parallel phrase pairs when presented with a large collection of phrase pair candidates. We evaluate the effectiveness of these approaches in detecting alignments for phrase pairs that have a known alignment in comparable sentence pairs. The results show that the Non-Viterbi alignment approach outperforms the other two approaches on F1 measure.

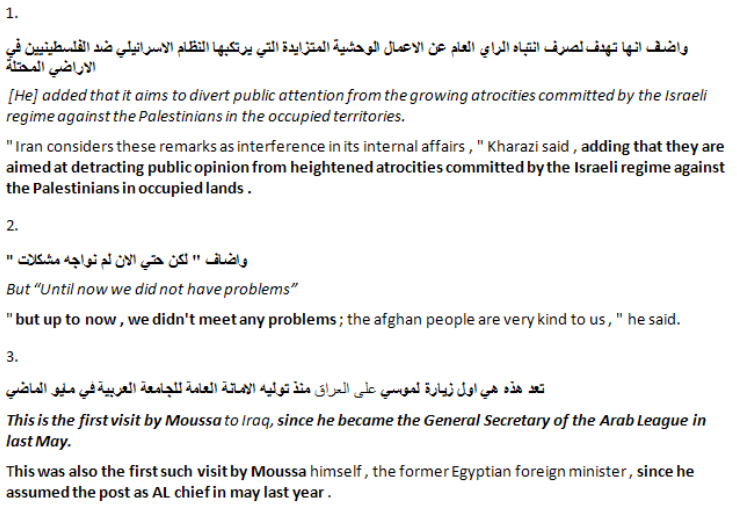
## 1 Introduction

Statistical Machine Translation (SMT), like many natural language processing tasks, relies primarily on parallel corpora. The translation performance of SMT systems directly depends on the quantity and the quality of the available parallel data. However, such corpora are only available in large quantities for a handful of languages, including English, Arabic, Chinese and some European languages. Much of this data is derived from parliamentary proceedings, though a limited amount of newswire text is also available. For most other languages, especially for less commonly used languages, parallel data is virtually non-existent.

Comparable corpora provide a possible solution to this data sparseness problem. Comparable documents are not strictly parallel, but contain rough translations of each other, with overlapping information. A good example for comparable documents is the newswire text produced by multilingual news organizations such as AFP or Reuters. The degree of parallelism can vary greatly, ranging from *noisy parallel* documents that contain many parallel sentences, to *quasi parallel* documents that may cover different topics (Fung and Cheung, 2004). The Web is by far the largest source of comparable data. Resnik and Smith (2003) exploit the similarities in URL structure, document structure and other clues for mining the Web for parallel documents. Wikipedia has become an attractive source of comparable documents in more recent work (Smith et al., 2010).

Comparable corpora may contain parallel data in different levels of granularity. This includes: parallel documents, parallel sentence pairs, or parallel sub-sentential fragments. To simplify the process and reduce the computational overhead, the parallel sentence extraction is typically divided into two tasks. First, a document level alignment is identified between comparable documents, and second, the parallel sentences are detected within the identified document pairs. Cross-lingual information retrieval methods (Munteanu and Marcu, 2005) and

*Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 61–68,
*49th Annual Meeting of the Association for Computational Linguistics*,
*Portland, Oregon, 24 June 2011.* ©*2011 Association for Computational Linguistics*



1.

**واضف انها تهدف لصرف انتباه الراي العام عن الاعمال الوحشية المتزايدة التي يرتكبها النظام الاسرائيلي ضد الفلسطينيين في الاراضي المحتلة**

*[He] added that it aims to divert public attention from the growing atrocities committed by the Israeli regime against the Palestinians in the occupied territories.*

" Iran considers these remarks as interference in its internal affairs , " Kharazi said , **adding that they are aimed at detracting public opinion from heightened atrocities committed by the Israeli regime against the Palestinians in occupied lands .**

2.

**واضاف " لكن حتي الان لم نواجه مشكلات "**

*But "Until now we did not have problems"*

" **but up to now , we didn't meet any problems** ; the afghan people are very kind to us , " he said.

3.

**تعد هذه هي اول زيارة لموسي** على العراق **منذ توليه الامانة العامة للجامعة العربية في مايو الماضي**

***This is the first visit by Moussa*** *to Iraq,* ***since he became the General Secretary of the Arab League in last May.***

**This was also the first such visit by Moussa** himself , the former Egyptian foreign minister , **since he assumed the post as AL chief in may last year .**

Figure 1: Sample comparable sentences that contain parallel phrases

other similarity measures (Fung and Cheung, 2004) have been used for the document alignment task. Zhao and Vogel (2002) have extended parallel sentence alignment algorithms to identify parallel sentence pairs within comparable news corpora. Tillmann and Xu (2009) introduced a system that performs both tasks in a single run without any document level pre-filtering. Such a system is useful when document level boundaries are not available in the comparable corpus.

Even if two comparable documents have few or no parallel sentence pairs, there could still be parallel sub-sentential fragments, including word translation pairs, named entities, and long phrase pairs. The ability to identify these pairs would create a valuable resource for SMT, especially for low-resource languages. The first attempt to detect sub-sentential fragments from comparable sentences is (Munteanu and Marcu, 2006). Quirk et al. (2007) later extended this work by proposing two generative models for comparable sentences and showed improvements when applied to cross-domain test data. In both these approaches the extracted fragment data was used as additional training data to train alignment models. Kumano et al. (2007) have proposed a phrasal alignment approach for comparable corpora using the joint probability SMT model. While this approach is appealing for low-resource scenarios as it does not require any seed parallel corpus, the high computational cost is a deterrent in its applicability to large corpora.

In this paper we explore several phrase alignment approaches to detect parallel phrase pairs embedded in comparable sentence pairs. We assume that comparable sentence pairs have already been detected. Our intention is to use the extracted phrases directly in the translation process, along with other phrase pairs extracted from parallel corpora. In particular, we study three alignment approaches:

- the standard phrase extraction algorithm, which relies on the Viterbi path of the word alignment;
- a phrase extraction approach that does not rely on the Viterbi path, but only uses lexical features;
- and a binary classifier to detect parallel phrase pairs when presented with a large collection of phrase pair candidates.

We evaluate the effectiveness of these approaches in detecting alignments for phrase pairs that have a known translation a comparable sentence pair. Section 2 introduces the phrase alignment problem in comparable sentences and discusses some of the challenges involved. It also explains the different alignment approaches we explore. Section 3 presents the experimental setup and the results of the evaluation. We conclude, in section 4, with an analysis of the results and some directions for future work.

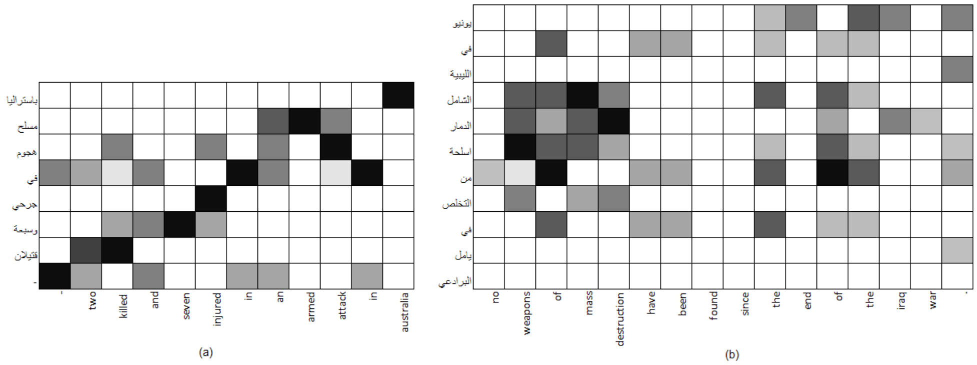


Figure 2: Word-to-word alignment pattern for (a) a parallel sentence pair (b) a non-parallel sentence pair

## 2 Parallel Phrase Extraction

Figure 1 shows three sample sentences that were extracted from Gigaword Arabic and Gigaword English collections. For each comparable sentence pair, the Arabic sentence is shown first, followed by its literal English translation (in Italics). The English sentence is shown next. The parallel sections in each sentence are marked in boldface. In the first two sentences pairs, the English sentence contains the full translation of the Arabic sentence, but there are additional phrases on the English side that are not present on the Arabic sentence. These phrases appear at the beginning of sentence 1 and at the end of sentence 2. In sentence 3, there are parallel phrases as well as phrases that appear only on one side. The phrase "to Iraq" appears only on the Arabic sentence while the phrase "the former Egyptian foreign minister" appears only on the English side.

Standard word alignment and phrase alignment algorithms are formulated to work on parallel sentence pairs. Therefore, these standard algorithms are not well suited to operate on partially parallel sentence pairs. Presence of non-parallel phrases may result in undesirable alignments.

Figure 2 illustrates this phenomenon. It compares a typical word alignment pattern in a parallel sentence pair (a) to one in a non-parallel sentence pair (b). The darkness of a square indicates the strength of the word alignment probability between the corresponding word pair. In 2(a), we observe high probability word-to-word alignments (dark squares) over the entire length of the sentences. In 2(b), we see one dark area above "weapons of mass destruction", corresponding to the parallel phrase pair, and some scattered dark spots, where high frequency English words pair with high frequency Arabic words. This spurious alignments pose problems to the phrase alignment, and indicate that word alignment probabilities alone might not be sufficient.

Our aim is to identify such parallel phrase pairs from comparable sentence pairs. In the following subsections we briefly explain the different phrase alignment approaches we use.

### 2.1 Viterbi Alignment

Here we use the typical phrase extraction approach used by Statistical Machine Translation systems: obtain word alignment models for both directions (source to target and target to source), combine the Viterbi paths using one of many heuristics, and extract phrase pairs from the combined alignment. We used Moses toolkit (Koehn et al., 2007) for this task. To obtain the word alignments for comparable sentence pairs, we performed a forced alignment using the trained models.

### 2.2 Binary Classifier

We used a Maximum Entropy classifier as our second approach to extract parallel phrase pairs from comparable sentences. Such classifiers have been used in the past to detect parallel sentence pairs in large collections of comparable documents (Munteanu and Marcu, 2005). Our classifier is similar, but we apply it at phrase level rather than at sentence level. The classifier probability is defined as:

$$p(c|S,T) = \frac{exp\left(\sum_{i=1}^{n} \lambda_i f_i(c,S,T)\right)}{Z(S,T)}, \quad (1)$$

where $S = s_1^L$ is a source phrase of length $L$ and $T = t_1^K$ is a target phrase of length $K$. $c \in \{0,1\}$ is a binary variable representing the two classes of phrases: *parallel* and *not parallel*. $p(c|S,T) \in [0,1]$ is the probability where a value $p(c=1|S,T)$ close to 1.0 indicates that $S$ and $T$ are translations of each other. $f_i(c,S,T)$ are feature functions that are co-indexed with respect to the class variable $c$. The parameters $\lambda_i$ are the weights for the feature functions obtained during training. $Z(S,T)$ is the normalization factor. In the feature vector for phrase pair $(S,T)$, each feature appears twice, once for each class $c \in \{0,1\}$.

The feature set we use is inspired by Munteanu and Marcu (2005) who define the features based on IBM Model-1 (Brown et al., 1993) alignments for source and target pairs. However, in our experiments, the features are computed primarily on IBM Model-1 probabilities (i.e. lexicon). We do not explicitly compute IBM Model-1 alignments. To compute coverage features, we identify alignment points for which IBM Model-1 probability is above a threshold. We produce two sets of features based on IBM Model-1 probabilities obtained by training in both directions. All the features have been normalized with respect to the source phrase length $L$ or the target phrase length $K$. We use the following 11 features:

1. Lexical probability (2): IBM Model-1 log probabilities $p(S|T)$ and $p(T|S)$

2. Phrase length ratio (2): source length ratio $K/L$ and target length ratio $L/K$

3. Phrase length difference (1): source length minus target length, $L - K$

4. Number of words covered (2): A source word $s$ is said to be covered if there is a target word $t \in T$ such that $p(s|t) > \epsilon$, where $\epsilon = 0.5$. Target word coverage is defined accordingly.

5. Number of words not covered (2): This is computed similarly to 4. above, but this time counting the number of positions that are not covered.

6. Length of the longest covered sequence of words (2)

To train the classifier, we used parallel phrases pairs extracted from a manually word-aligned corpus. In selecting negative examples, we followed the same approach as in (Munteanu and Marcu, 2005): pairing all source phrases with all target phrases, but filter out the parallel pairs and those that have high length difference or a low lexical overlap, and then randomly select a subset of phrase pairs as the negative training set. The model parameters are estimated using the GIS algorithm.

### 2.3 Non-Viterbi (PESA) Alignment

A phrase alignment algorithm called “PESA” that does not rely on the Viterbi path is described in (Vogel, 2005). PESA identifies the boundaries of the target phrase by aligning words inside the source phrase with words inside the target phrase, and similarly for the words outside the boundaries of the phrase pair. It does not attempt to generate phrase alignments for the full sentence. Rather, it identifies the best target phrase that matches a given source phrase. PESA requires a statistical word-to-word lexicon. A seed parallel corpus is required to automatically build this lexicon.

This algorithm seems particularly well suited in extracting phrase pairs from comparable sentence pairs, as it is designed to not generate a complete word alignment for the entire sentences, but to find only the target side for a phrase embedded in the sentence. We briefly explain the PESA alignment approach below.

Instead of searching for all possible phrase alignments in a parallel sentence pair, this approach finds the alignment for a single source phrase $S = s_1 \dots s_l$. Assume that we have a parallel sentence pair $(s_1^J, t_1^I)$ which contains the source phrase $S$ in the source sentence $s_1^J$. Now we want to find the target phrase $T = t_1 \dots t_k$ in the target sentence $t_1^I$ which is the translation of the source phrase.
A constrained IBM Model-1 alignment is now applied as follows:

- Source words inside phrase boundary are aligned only with the target words inside the phrase boundary. Source words outside the phrase boundary are only aligned with target words outside the phrase boundary.

- Position alignment probability for the sentence, which is $1/I$ in IBM Model-1, is modified to be $1/k$ inside the source phrase and to $1/(I-k)$ outside the phrase.

Figure 3 shows the different regions. Given the source sentence and the source phrase from position $j_1$ to $j_2$, we want to find the boundaries of the target phrase, $i_1$ and $i_2$. The dark area in the middle is the phrase we want to align. The size of the blobs in each box indicates the lexical strength of the word pair.

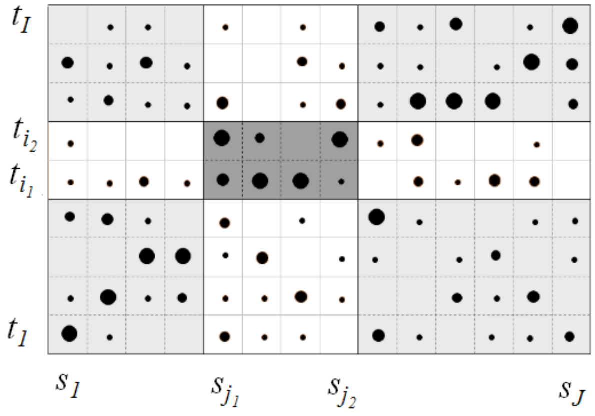


Figure 3: PESA Phrase alignment

The constrained alignment probability is calculated as follows:

$$\begin{aligned} p(s|t) &= \left( \prod_{j=1}^{j_1-1} \sum_{i \notin (i_1 \ldots i_2)} \frac{1}{I-k} p(s_j|t_i) \right) \\ &\times \left( \prod_{j=j_1}^{j_2} \sum_{i=i_1}^{i_2} \frac{1}{k} p(s_j|t_i) \right) \qquad (2) \\ &\times \left( \prod_{j=j_2+1}^{J} \sum_{i \notin (i_1 \ldots i_2)} \frac{1}{I-k} p(s_j|t_i) \right) \end{aligned}$$

$p(t|s)$ is similarly calculated by switching source and target sides in equation 2:

$$\begin{aligned} p(t|s) &= \left( \prod_{i=1}^{i_1-1} \sum_{i \notin (j_1 \ldots j_2)} \frac{1}{J-l} p(t_i|s_j) \right) \\ &\times \left( \prod_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{1}{l} p(t_i|s_j) \right) \qquad (3) \\ &\times \left( \prod_{i=i_2+1}^{I} \sum_{j \notin (j_1 \ldots j_2)} \frac{1}{J-l} p(t_i|s_j) \right) \end{aligned}$$

To find the optimal target phrase boundaries, we interpolate the two probabilities in equations 2 and 3 and select the boundary $(i_1, i_2)$ that gives the highest probability.

$$\begin{aligned} (i_1, i_2) = \underset{i_1, i_2}{\operatorname{argmax}} \; & \{(1-\lambda)\, log(p(s|t)) \\ & + \lambda\, log(p(t|s))\} \qquad (4) \end{aligned}$$

The value of $\lambda$ is estimated using held-out data.

PESA can be used to identify all possible phrase pairs in a given parallel sentence pair by iterating over every source phrase. An important difference is that each phrase is found independently of any other phrase pair, whereas in the standard phrase extraction they are tied through the word alignment of the sentence pair.

There are several ways we can adapt the non-Viterbi phrase extraction to comparable sentence.

- Apply the same approach assuming the sentence pair as parallel. The inside of the source phrase is aligned to the inside of the target phrase, and the outside, which can be non-parallel, is aligned the same way.

- Disregard the words that are outside the phrase we are interested in. Find the best target phrase by aligning only the inside of the phrase. This will considerably speed-up the alignment process.

## 3 Experimental Results

### 3.1 Evaluation Setup

We want to compare the performance of the different phrase alignment methods in identifying parallel phrases embedded in comparable sentence pairs.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | All |
|---|---|---|---|---|---|---|---|---|---|---|---|
| test set | 2,826 | 3,665 | 3,447 | 3,048 | 2,718 | 2,414 | 2,076 | 1,759 | 1,527 | 1,378 | 24,858 |
| test set (found) | 2,746 | 2,655 | 1,168 | 373 | 87 | 29 | 7 | 2 | 1 | 0 | 7,068 |

Table 1: N-gram type distribution of manually aligned phrases set

Using a manually aligned parallel corpus, and two monolingual corpora, we obtained a test corpus as follows: From the manually aligned corpus, we obtain parallel phrase pairs $(S, T)$. Given a source language corpus $\mathcal{S}$ and a target language corpus $\mathcal{T}$, for each parallel phrase pair $(S, T)$ we select a sentence $s$ from $\mathcal{S}$ which contains $S$ and a target sentence $t$ from $\mathcal{T}$ which contains $T$. These sentence pairs are then non-parallel, but contain parallel phrases, and for each sentence pair the correct phrase pair is known. This makes it easy to evaluate different phrase alignment algorithms.

Ideally, we would like to see the correct target phrase $T$ extracted for a source phrase $S$. However, even if the boundaries of the target phrase do not match exactly, and only a partially correct translation is generated, this could still be useful to improve translation quality. We therefore will evaluate the phrase pair extraction from non-parallel sentence pairs also in terms of partial matches.

To give credit to partial matches, we define precision and recall as follows: Let $W$ and $G$ denote the extracted target phrase and the correct reference phrase, respectively. Let $M$ denote the tokens in $W$ that are also found in the reference $G$. Then

$$Precision = \frac{|M|}{|W|} * 100 \tag{5}$$

$$Recall = \frac{|M|}{|G|} * 100 \tag{6}$$

These scores are computed for each extracted phrase pair, and are averaged to produce precision and recall for the complete test set. Finally, precision and recall are combined to generated the F-1 score in the standard way:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{7}$$

### 3.2 Evaluation

We conducted our experiments on Arabic-English language pair. We obtained manual alignments for 663 Arabic-English sentence pairs. From this, we selected 300 sentences, and extracted phrase pairs up to 10 words long that are consistent with the underlying word alignment. From the resulting list of phrase pairs, we removed the 50 most frequently occurring pairs as well as those only consisting of punctuations. Almost all high frequency phrases are function words, which are typically covered by the translation lexicon. Line 1 in Table 1 gives the n-gram type distribution for the source phrases.

Using the phrase pairs extracted from the manually aligned sentences, we constructed a comparable corpus as follows:

1. For each Arabic phrase, we search the Arabic Gigaword[1] corpus for sentences that contain the phrase and select up to 5 sentences. Similarly, for each corresponding English phrase we select up to 5 sentences from English Gigaword[2].

2. For each phrase pair, we generate the Cartesian product of the sentences and produce a sentence pair collection. I.e. up to 25 comparable sentence pairs were constructed for each phrase pair.

3. We only select sentences up to 100 words long, resulting in a final comparable corpus consisting of 170K sentence pairs.

Line 2 in Table 1 gives the n-gram type distribution for the phrase pairs for which we found both a source sentence and a target sentence in the monolingual corpora. As expected, the longer the phrases, the less likely it is to find them in even larger corpora.

We consider the resulting set as our comparable corpus which we will use to evaluate all alignment approaches. In most sentence pairs, except for the phrase pair that we are interested in, the rest of the sentence does not typically match the other side.

[1] Arabic Gigaword Fourth Edition (LDC2009T30)

[2] English Gigaword Fourth Edition (LDC2009T13)

| Lexicon | Viterbi | | | | Classifier | | | | PESA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Exact | P | R | F1 | Exact | P | R | F1 | Exact | P | R | F1 |
| Lex-Full | 43.56 | 65.71 | 57.99 | 61.61 | 54.46 | 81.79 | 85.29 | 85.29 | 67.94 | 93.34 | 86.80 | 90.22 |
| Lex-1/3 | 42.95 | 65.68 | 56.69 | 60.85 | 53.57 | 81.32 | 88.34 | 84.69 | 67.28 | 93.23 | 86.17 | 89.56 |
| Lex-1/9 | 41.10 | 63.60 | 51.15 | 56.70 | 52.38 | 80.30 | 86.64 | 83.35 | 65.81 | 91.95 | 84.73 | 88.19 |
| Lex-1/27 | 41.02 | 62.10 | 49.38 | 55.01 | 52.51 | 80.51 | 83.84 | 82.14 | 63.23 | 89.41 | 82.06 | 85.57 |
| Lex-BTEC | 19.10 | 26.94 | 23.63 | 25.18 | 18.76 | 45.90 | 36.17 | 40.46 | 17.45 | 46.70 | 36.28 | 40.83 |

Table 2: Results for Alignment Evaluation of test phrases

We obtained the Viterbi alignment using standard word alignment techniques: IBM4 word alignment for both directions, Viterbi path combination using heuristics ('grow-diag-final') and phrase extraction from two-sided training, as implemented in the Moses package (Koehn et al., 2007). Because the non-parallel segments will lead the word alignment astray, this may have a negative effect on the alignment in the parallel sections. Alignment models trained on parallel data are used to generate the Viterbi alignment for the comparable sentences. We then extract the target phrases that are aligned to the embedded source phrases. A phrase pair is extracted only when the alignment does not conflict with other word alignments in the sentence pair. The alignments are not constrained to produce contiguous phrases. We allow unaligned words to be present in the phrase pair. For each source phrase we selected the target phrase that has the least number of unaligned words.

The classifier is applied at the phrase level. We generate the phrase pair candidates as follows: For a given target sentence we generate all n-grams up to length 10. We pair each n-gram with the source phrase embedded in the corresponding source sentence to generate a phrase pair. From the 170 thousand sentence pairs, we obtained 15.6 million phrase pair candidates. The maximum entropy classifier is then applied to the phrase pairs. For each source phrase, we pick the target candidate for which $p(c = 1, S, T)$ has the highest value.

For the PESA alignment we used both inside and outside alignments, using only lexical probabilities. For each source phrase pair, we select the best scoring target phrase.

As our goal is to use these methods to extract parallel data for low resource situations, we tested each method with several lexica, trained on different amounts of initial parallel data. Starting from the full corpus with 127 million English tokens, we generated three additional parallel corpora with 1/3, 1/9 and 1/27 of the original size. The 1/9 and 1/27 corpora (with 13 million and 4 million English words) can be considered *medium* and *small* sized corpora, respectively. These two corpora are a better match to the resource levels for many languages. We also used data from the BTEC (Kikui et al., 2003) corpus. This corpus contains conversational data from the travel domain, which is from a different genre than the document collections. Compared to other corpora, it is much smaller (about 190 thousand English tokens).

Table 2 gives the results for all three alignment approaches. Results are presented as percentages of: exact matches found (Exact), precision (P), recall (R) and F1. The Viterbi alignment gives the lowest performance. This shows that the standard phrase extraction procedure, which works well for parallel sentence, is ill-suited for partially parallel sentences. Despite the fact that the classifier incorporates several features including the lexical features, the performance of the PESA alignment, which uses only the lexical features, has consistently higher precision and recall than the classifier. This demonstrates that computing both inside and outside probabilities for the sentence pair helps the phrase extraction. The classifier lacks this ability because the phrase pair is evaluated in isolation, without the context of the sentence.

Except for the BTEC corpus, the performance degradation is minimal as the lexicon size is reduced. This shows that the approaches are robust for smaller parallel amounts of parallel data.

Instead of using token precision, an alternative method of evaluating partial matches, is to give credit based on the length of the overlap between the extracted phrase and the reference. Precision and recall can then be defined based on the longest common contiguous subsequence, similar to (Bourdaillet et al., 2010). Results obtained using this methods were similar to the results in Table 2.

## 4 Conclusion and Future Work

In this paper we explored several phrase alignment approaches for extracting phrase pairs that are embedded inside comparable sentence pairs. We used the standard Viterbi phrase alignment, a maximum entropy classifier that works on phrase pairs, and a non-Viterbi PESA alignment in the evaluation process. The results show that PESA outperforms both the Viterbi approach and the classifier, in both precision and recall.

We plan to extend the PESA framework to use not only lexical features, but other features similar to the ones used in the classifier. We believe this will further improve the alignment accuracy.

While this paper focuses on comparisons of different phrase alignment approaches in a realistic, yet controlled manner by selecting appropriate comparable sentence pairs for given phrase pairs, future experiments will focus on finding new phrase pairs from comparable corpora and evaluating the potential utility of the extracted data in the context of an end-to-end machine translation system.

## References

Julien Bourdaillet, Stéphane Huet, Philippe Langlais, and Guy Lapalme. 2010. TransSearch: from a bilingual concordancer to a translation finder. *Machine Translation*, 24(3-4):241–271, dec.

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

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

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *In Proc. of EUROSPEECH 2003*, pages 381–384, Geneva.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June.

Tadashi Kumano, Hideki Tanaka, and Takenobu Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. In *In Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation*, Skvde, Sweden, September.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia.

Chris Quirk, Raghavendra U. Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of the Machine Translation Summit XI*, pages 377–384, Copenhagen, Denmark.

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

Jason R. Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Proceedings of the Human Language Technologies/North American Association for Computational Linguistics*, pages 403–411.

Christoph Tillmann and Jian-Ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Companion Vol. of NAACL HLT 09*, Boulder, CA, June.

Stephan Vogel. 2005. PESA: Phrase pair extraction as sentence splitting. In *Proceedings of the Machine Translation Summit X*, Phuket, Thailand, September.

Bing Zhao and Stephan Vogel. 2002. Full-text story alignment models for chinese-english bilingual news corpora. In *Proceedings of the ICSLP '02*, September.