Revisiting sentence alignment algorithms for alignment visualization and evaluation

Qian Yu, Aurélien Max, François Yvon

LIMSI-CNRS and Univ. Paris Sud
rue John Von Neuman, F-91 403 Orsay
yu@limsi.fr, amax@limsi.fr, yvon@limsi.fr

Abstract

In this paper, we revisit the well-known problem of sentence alignment, in a context where the entire bitext has to be aligned and where alignment confidence measures have to be computed. Following much recent work, we study here a multi-pass approach: we first compute sure alignments that are used to train a discriminative model; then we use this model to fill in the gaps between sure links. Experimental results on several corpora show the effectiveness of this method as compared to alternative, state-of-the-art, proposals.

1. Introduction

The alignment of *bitexts*, *i.e.* of pairs of texts assumed to be mutual translations, consists in finding correspondences between logical units in parallel texts. The set of such correspondences is called an *alignment*. Depending on the logical units that are considered, various levels of granularity for the alignment are obtained. It is for usual to compute alignments at the level of paragraphs, sentences, phrases or words (see (Wu, 2010; Tiedemann, 2011) for two recent reviews). Alignments are widely used in many fields, especially in multilingual text processing (multilingual Information Retrieval, multilingual terminology extraction and Machine Translation). For all these applications, alignments between sentences must be computed.

Sentence alignment is generally considered an easy task and many sentence alignment algorithms have been proposed in the literature. From a bird's eye view, two main families of approaches can be isolated, which both rely on the assumption that the relative order of sentences is the same on the two sides of the bitext. On the one hand, *length-based approaches* (Gale and Church, 1991; Brown et al., 1991) use the fact that the translation of a short (resp. long) sentence is short (resp. long). On the other hand, *lexical matching approaches* (Kay and Röscheisen, 1993; Simard et al., 1993) identify sure anchor points for the alignment using bilingual dictionaries or surface similarities of word forms. Length-based approaches are fast but error-prone, while lexical matching approaches seem to deliver more reliable results. Most recent, state-of-the art approaches to the problem (Langlais, 1998; Simard and Plamondon, 1998; Moore, 2002; Braune and Fraser, 2010) try to combine both types of information.

In most applications, notably for training Machine Translation systems, only high-confidence, one-to-one, sentence alignments are kept. Indeed, when the objective is to build subsentential (phrase or word) alignments, the other types of mappings between sentences are deemed to be either insufficiently reliable or inappropriate. As it were, the one-to-one constraint is viewed as a proxy to literalness/compositionality of the translation, and warrants the search for finer-grained alignments. However, for certain types of bitexts, for instance literary texts, translation often departs from a straight sentence-by-sentence mode and using such a constraint discards a significant portion of the bitext. For Machine Translation, this is just a regrettable waste of potential training material. For other applications, however, notably applications which imply to visualize or read the actual translations in their context, as is, for instance, the case for second language learning, for training translators, or for automatic translation checking (Macklovitch, 1994), the entire bitext has to be aligned. Furthermore, areas where the translation is only partial or approximative may have to be identified precisely.

Following much recent work, we explore here a multiple-pass approach to sentence alignment. In a nutshell, our approach relies on sure one-to-one mappings detected in a first pass to train a discriminative sentence alignment system, which is then used to align the regions which remain problematic. Our experiments on the BAF corpus (Simard, 1998) show that this approach produces very high quality alignments, and also allows to identify the most problematic passages.

The rest of this paper is organized as follows: we first briefly review existing alignment methods in Section 2. In Section 3., we evaluate these methods and analyze the main sentence alignment errors. Our algorithm is detailed in Section 4., and evaluated on standard benchmarks in Section 5. We discuss further prospects and conclude in Section 6.

2. Sentence alignment: a review

Sentence alignment is an old task and the first proposals date back more than twenty years ago. These initial attempts can roughly be classified in two main categories: *length-based approaches* and *lexical matching approaches* (Tiedemann, 2011). The former family of approaches are based on the correlation of the length of parallel sentences, as introduced independently by Gale and Church (1991) and by Brown et al. (1991). The main intuition here is that long source sentences align preferably with long target sentences, and short source sentences with short target sentences. The difference between these two proposals is the way length is measured: the former study uses the number of characters, while the latter uses the number of words. The second family of approaches rely on sure or obvious lexical correspondences, as provided, for instance, by entries of a bilingual dictionary, by so-called orthographical cognates[1] (Simard et al., 1993), or by word pairs having similar distributions of occurrence (Kay and Röscheisen, 1993). In both cases, additional simplifying assumptions are used, notably the fact that the relative order of sentences is preserved, and that sentences mostly align near the "diagonal" of the bitext, thus yielding very efficient algorithms.

Realizing the shortcomings of these initial proposals, several authors have proposed ways to combine the length-based approach and the lexical matching approach for aligning sentences (Chen, 1993; Wu, 1994; Moore, 2002; Braune and Fraser, 2010). For instance, the method proposed by Moore (2002) uses a three-step process for aligning sentences. First, a coarse alignment of the corpus is computed using a modified version of Brown et al.'s length-based model where search pruning techniques are used to speed up the discovery of reliable sentence pairs. In a second stage, the sentence pairs having the highest alignment probability are collected to train a modified version of IBM Translation Model 1 (Brown et al., 1993). Finally, the entire corpus is realigned using the IBM Model 1 score as an additional measure of parallelism. This method achieves high accuracy at a modest computational cost and does not require any knowledge of the languages or the corpus except how to break up the text into words and sentences. A very similar multi-pass approach is proposed in (Braune and Fraser, 2010), which basically aims at improving the unsatisfactory recall of Moore's algorithm, which misses many matchings when the bitext are not completely parallel.

Recent years have witnessed very few new proposals for this task and the problem seems to be basically solved. The only notable exceptions are the work of Deng et al. (2007), which tries to go beyond one-to-one sentence alignments, and considers matching large subparts using a divisive segmentation algorithm; the work of Fattah et al. (2007) using supervised learning tools; the robust aligner of Ma (2006), which relies on a statistical weighting scheme to balance the significance of bilingual lexical matches in parallel sentences; and the study of Sennrich and Volk (2010), which considers monolingual sentence alignment techniques after automatically projecting target texts back to the source language with machine translation.

3. A systematic analysis of alignment errors

3.1. Corpus and Baselines

In a first attempt to evaluate existing alignment methods, we selected a French literary work "De la terre à la Lune" by Jules Verne and its English translation "From the earth to the moon". This book is available as part of the BAF corpus (Simard, 1998). The French side of the bitext contains 3,319 sentences, 69,456 running words and 347,691 characters, whereas the English version contains 2,554 sentences, 50,331 words and 245,657 characters. Note the large difference in length between the French and the English side: as previously noted, the translation is only approximative, and it often appears that French paragraphs are summarized, rather than translated, into one or two English sentences. Both texts are shipped with reference sentence segmentations and alignment links.

To make our experimentations easier, we used the Uplug package[2], which provides a unified interface to integrate various sentence alignment methods. The distribution of Uplug ships with several alignment algorithms: the Gale-Church method[3], GMA[4] (Melamed, 1999), hunalign[5] (Varga et al., 2005), and some others. To these, we added the Moore aligner[6], the Gargantua alignment system[7] and BleuAlign[8]. All the input and output files are in the same format, which makes experimentation and inter-system comparison much easier.

Results are given in Table 1, where we display recall, precision and F-measure computed *at the alignment and at the sentence level*[9]. Note that with the latter metric, errors on $0 - n$ or $n - 0$ alignments are not taken into account. This might be because[10] it is generally considered unimportant to miss such alignments, which are not useful in the perspective of building parallel training material for Machine Translation. As reported in this table, some methods have very good precision, while recall is on average less satisfactory; the most extreme case is Moore's method, which achieves a nearly perfect precision, at the expense of a much worse recall.

3.2. Error analysis

As previously noted by several authors, this corpus is difficult because of the relatively low proportion of 1-to-1 links. This may be due to the use of non-literal translations or to differences in sentence segmentation. As detailed in Table 2, most methods are unable to reproduce the reference link distribution. The main issue is with null links, which, in this corpus, account for approximately 8.6% of the alignments. Only GMA is getting close to the right distribution, at the price, though, of a precision less satisfactory than for other approaches. It should be noted that making errors on such links often cause the desynchronization of entire passages, which has a strong negative impact on performance.

[1] Cognates are words that share a similar spelling in two or more different languages, as a result of their similar meaning and/or common etymological origin, e.g. (English-Spanish): history - historia, harmonious - armonioso. In subsequent references, they are more loosely defined as two words in different languages sharing a common prefix.

[2] http://sourceforge.net/projects/uplug/

[3] Using the implementation of Michael D. Riley.

[4] http://nlp.cs.nyu.edu/GMA/

[5] ftp://ftp.mokk.bme.hu/Hunglish/src/hunalign

[6] http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/

[7] http://sourceforge.net/projects/gargantua/

[8] https://github.com/rsennrich/bleualign/

[9] Other useful metrics for sentence alignment are based on recall and precision computed at the level of words and characters (see e.g. (Véronis and Langlais, 2000)).

[10] P. Langlais, personal communication.

	Gale	GMA	Hunalign	Moore	Gargantua
Alignment based metrics					
precision	0.30	0.61	0.50	0.85	0.74
recall	0.29	0.65	0.59	0.65	0.71
F-measure	0.29	0.63	0.54	0.74	0.72
Sentence based metrics					
precision	0.34	0.75	0.74	0.98	0.88
recall	0.39	0.77	0.69	0.62	0.77
F-measure	0.36	0.76	0.71	0.76	0.82

Table 1: Performance of various sentence alignment algorithms

Link type	0–1:5	1:5–0	1–1	1–2:5	2–1	2–2:5	others
Reference	0.56	8.05	75.71	4.37	4.60	3.65	3.06
Gale	0	0.41	59.22	3.51	34.63	2.23	0
GMA	0.74	10.54	68.42	4.43	13.02	1.00	1.85
Hunalign	0.20	1.41	61.02	3.93	33.44	0	0
Moore	0	0	100	0	0	0	0
Gargantua	0	0	91.64	3.97	3.85	0	0.54

Table 2: Distribution of predicted alignment types
Column 0-1:5 *gathers all the alignments matching* 0 *source sentence with* $1 \leq n \leq 5$ *target sentences.*

4. A coarse-to-fine approach to sentence alignment

4.1. Overview

We introduce in this section our coarse-to-fine alignment strategy. As with most multi-pass approaches, the first step is meant to provide a computationally cheap way to drastically reduce the alignment search space, by providing us with a first set of very high precision alignment links. All of these sentences that are aligned during this step are used as anchor points for the second step; they are also used to train a classifier aimed at recognizing parallel groups of sentences. The second step of our method uses an exhaustive search to enumerate and evaluate all the possible ways to align the blocks that appear between two anchor points. Based on the previous analysis, these blocks are typically sufficiently small that an exhaustive search is actually feasible. Based on these evaluations, a greedy algorithm is finally used to select the sentence pairs that align with highest probabilities.

For the first step, we simply chose to use the method of Moore (2002) because of its excellent precision. A tighter integration between this first step and the subsequent computations, which require to recompute several statistics that are used in Moore's approach, is certainly desirable. Yet, at this stage, we favored simplicity over computational efficiency. The two other steps are detailed below.

4.2. Detecting parallelism

The second step of our approach consists in training a function for scoring candidate alignments. Following (Munteanu and Marcu, 2005), we used a Maximum Entropy model[11] (Rathnaparkhi, 1998); in principle, many other choices would be possible here. We take the sentence alignments of the first step as positive examples; for negative examples, we randomly chose pairs $(\mathbf{e}, \mathbf{f}')$, where $(\mathbf{e}, \mathbf{f})$ and $(\mathbf{e}', \mathbf{f}')$ are two positive instances and $\mathbf{e}'$ directly follows $\mathbf{e}$. This strategy produced a balanced corpus containing as many negative pairs as positive ones. However, this approach may give too much weight on the length ratio feature and it remains to be seen whether alternative approaches are more suitable.

Our problem is thus to estimate a conditional model for deciding whether two sentences $\mathbf{e}$ and $\mathbf{f}$ should be aligned. Denoting Y the corresponding binary variable, this model has the following form:

$$P(Y = 1|\mathbf{e}, \mathbf{f}) = \frac{1}{1 + exp[-\sum_{i=1}^{k} \theta_k F_k(\mathbf{e}, \mathbf{f})]},$$

where $\{F_k(\mathbf{e}, \mathbf{f}), k = 1 \ldots K\}$ denotes a set of feature functions testing arbitrary properties of $\mathbf{e}$ and $\mathbf{f}$ and $\{\theta_k, k = 1 \ldots K\}$ is the corresponding set of parameter values.

Given a set of training sentence pairs, the optimal values of the parameters are set by optimizing numerically the conditional likelihood; optimization is performed here using L-BFGS (Liu and Nocedal, 1989); a Gaussian prior over the parameters is used to ensure numerical stability of the optimization. In practice, this means that the objective function is the inverse of the conditional log-likelihood, completed with a quadradic term proportional to $\sum_{i=1}^{k} \theta^2$.

In this study, we used the following set of feature functions:

- **lexical features**: for each pair of words[12] (e, f) occurring in $V_e \times V_f$, there is a corresponding feature $F_{e,f}$ which fires whenever $e \in \mathbf{e}$ and $f \in \mathbf{f}$.
- **length features**: denoting $l_{\mathbf{e}}$ (resp. $l_{\mathbf{f}}$) the length of the source (resp. target) sentence, measured in num-

[11] We use the implementation available from http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

[12] A word is an alphabetic string of characters, excluding punction marks.

ber of characters, we include features related to length ratio, defined as $F_r(\mathbf{e}, \mathbf{f}) = \frac{|l_\mathbf{e} - l_\mathbf{f}|}{max(l_\mathbf{e}, l_\mathbf{f})}$. Rather than taking the numerical value, we use a simple discretization scheme based on 6 bins.

4.3. Filling alignment gaps

The third step uses the posterior alignment probabilities computed in the second step to fill the gaps in the first pass alignments. The algorithm can be glossed as follows. Assume a bitext block comprising the sentences from index i to j in the source side of the bitext, and from k to l in the target side such that sentences $\mathbf{e}_{i-1}$ (resp. $\mathbf{e}_{j+1}$) and $\mathbf{f}_{k-1}$ (resp. $\mathbf{e}_{l+1}$) are aligned[13].

The first case is when $j < i$ or $k > l$, in which case we create a null alignment for $f_{k:l}$ or for $e_{i:j}$. In all other situations, we compute:

$$\forall i', j', k', l', i \leq i' \leq j' \leq j, k \leq k' \leq l' \leq l,$$
$$a_{i',j',k',l'} = P(Y = 1 | \mathbf{e}_{i':j'}, \mathbf{f}_{k':l'}),$$

where $\mathbf{e}_{i':j'}$ is obtained by concatenation of all the sentences in the range $i' : j'$. Note that this implies to compute $O(|j - i|^2 \times |k - l|^2)$ probabilities, which, given the typical size of these blocks (see below), can be performed very quickly.

These values are then iteratively visited by decreasing order in a greedy fashion. The top-scoring block $i' : j', k' : l'$ of the list is retained in the final alignment; all blocks that overlap with this block are deleted from the list and the next best entry is then considered. This process continues until all remaining blocks imply null alignments, in which case these $n - 0$ or $0 - n$ alignments are also included in our solution.

This process is illustrated on Figure 1: assuming that the best matching link is f_2-e_2, we delete all the links that include f_2 or e_2, as well as links that would imply a reordering of sentences, meaning that we also delete links such as f_1-e_3 etc.

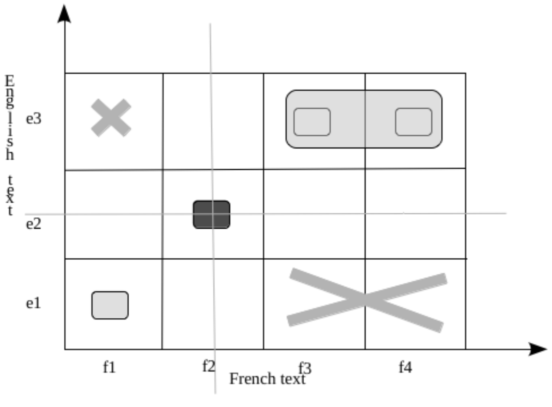

Figure 1: Greedy alignment search

5. Experiments

5.1. Results on literary work

In this first round of experiments, we consider the same literary work as in section 3.

The first alignment step, using Moore's algorithm with default parameters, identifies 1936 one-to-one alignments, used as anchor points for the remaining steps of the procedure. These are high-quality alignments which only contain 35 errors. Nonetheless, this first step creates a small number of misalignments: these errors can not be fixed, introduce some noise in the training set of the classifier and will also create more alignment errors in the subsequent steps. Using these anchor points, 447 "paragraphs" need to be further processed, corresponding to 1,383 French and 618 English sentences: the average length of these paragraphs is then respectively 2.9 sentences for the French side and 1.3 sentences for the English side, which makes our search procedure for fine-grained alignments computationally tractable. Note that not all these paragraphs need to be processed: in fact, for 156 of them, the only possible decision is to align 0-to-n or n-to-0.

In order to assess the quality of the Maxent classifier, we split the available training data into 90% for estimating the parameters and 10% for testing, and found that its decisions were correct 75% of the time[14].

A contrast was run using a much larger corpus of parallel sentences extracted from a collection of literary bitext. Here, the total number of training sentences was 133,562. Increasing the number of training sentences increased the precision of the classifier from 75% to 81%.

The third step was to fill the alignment gap using the algorithm presented in previous section. Here again, two strategies were tested: the baseline approach is a faithful implementation of our approach; alternatively, we tried to discard all the alignment links whose probability (for the Maxent model) is less than 0.5. In this condition, the number of null alignments is significantly increased.

The results of our experiments are summarized in Table 3. As reflected by these results, our multi-pass strategy delivers alignment results that significantly improve over the state-of-the-art. Unsurprisingly, we were able to boost the initial recall of Moore's method at the cost of only a small lost in precision. The F-measure is better than all the other alignment techniques, slightly surpassing the recent proposal of Braune and Fraser (2010). Using a larger training corpus has a small effect on the precision of the Maxent classifier, which does not show on the global alignment performance: our classifier is arguably delivering better performance, but its feature weights are less adapted to the specificities of our data. Likewise, using a prefiltering stage has hardly any impact on the global quality of our results; yet, this filtering is useful for speeding up our algorithm as it enables to discard 93% of the potential alignment links.

Looking at errors by alignment types (Table 4), we see that our method is able to better reproduce the distribution of link types, even though 0-to-n links still account for a substantial number of errors.

A qualitative analysis of alignment errors showed that:

[13] We conventionally enclose the source and target texts between begin and end markers so as to ensure that the first and last sentences are aligned.

[14] These results were obtained using 10-fold cross validation.

Score	Moore+Maxent	Moore+Maxent (+corpus)	Moore+Maxent (filtering_0.5)
Alignment based metrics			
precision	0.74	0.74	0.72
recall	0.81	0.80	0.80
F-measure	0.77	0.77	0.76
Sentence based metrics			
precision	0.93	0.90	0.94
recall	0.80	0.80	0.78
F-measure	0.86	0.85	0.85

Table 3: Performance at the alignment level and sentence level

	Link type						
	0–1:5	1:5–0	1–1	1–2:5	2–1	2–2:5	others
Reference	0.56	8.05	75.71	4.37	4.60	3.65	3.06
Moore+Maxent	2.64	10.73	79.88	2.13	2.46	0.44	1.72

Table 4: Distribution of predicted link types

- modeling null alignments remains difficult, as these links are only produced as a fall-back decision, for lack of finding better alignments. As a result, these alignments continue to account for a large number of errors.
- the model we train to predict alignment probability is a "bag-of-words" model and is only concerned with the cooccurrence of words in the French and English side, no matter how often these words occur. As a result, two adjacent sentences using the same vocabulary tend to confuse our aligners. This also occurs when adjacent sentences contain word pairs that were not seen in training and which play no role in scoring the alignments: the system is then unable to choose between segmenting a block of sentences or keeping them as a group (see examples in Figure 2).

A last question concerns the use of the model's scores as confidence estimation measures for the alignment. To check this, we removed from the final alignment all the blocks whose score is below a given threshold $0 \leq \theta \leq 1$ for varying values of θ; by convention, we assume that Moore's alignment links are sure and are never discarded. As expected, increasing θ from 0 (no filtering) to 1 (filter all but Moore's blocks) increases the precision, but is detrimental to recall. A slightly better F-measure of 0.78 can be obtained for $\theta = 0.4$; the variations are however small and remain to be confirmed for larger scale studies.

5.2. Complementary results on the BAF

In this section, we report on experiments conducted with other documents contained in the BAF corpus. Our goal here is to check that our method, which performs quite well on a "difficult" text, is also able to handle the easier types, such as institutional texts or scientific articles[15]. Our results are summarized in Table 5, where we compare our approach with its main competitors and show that it attains state-of-the-art results on these collections as reflected by the comparison with the Gargantua software.

	Moore	Gargantua	Moore+Maxent
Institutional texts			
Alignment based metrics			
precision	0.97	0.96	0.93
recall	0.91	0.96	0.95
F-measure	0.94	0.96	0.94
Sentence based metrics			
precision	0.99	0.98	0.98
recall	0.84	0.93	0.93
F-measure	0.91	0.95	0.95
Scientific articles			
Alignment based metrics			
precision	0.89	0.86	0.85
recall	0.89	0.91	0.93
F-measure	0.89	0.88	0.89
Sentence based metrics			
precision	1.00	0.98	0.95
recall	0.72	0.77	0.81
F-measure	0.84	0.86	0.87

Table 5: Performance at the alignment level and sentence level on other parts of the BAF corpus

6. Conclusions

In this paper, we have presented a novel two-pass approach aimed at improving existing sentence alignment methods in contexts where (i) all sentences need to be aligned and/or (ii) sentence alignment confidence need to be computed. By running experiments with several variants of this approach, we have been able to show that it was slightly better than the state-of-the-art on aligning a novel with its translation, and equivalent to the best approaches on other benchmarks. These results will be complemented by our ongoing experiments with the other benchmarks of Arcade 2 (Chiao et al., 2006) and with other literary corpora.

[15] As is standard pratice, we have not tried to align the technical manuals, which pose specific and difficult alignment problems.

(src) =”1.2065”	un second tiers voyait mal et n’ entendait pas ;
(src) =”1.2066”	quant au troisième , il ne voyait rien et n’ entendait pas davantage .
(trg) =”1.1555”	a second set saw badly and heard nothing at all ;
(trg) =”1.1556”	and as for the third , it could neither see nor hear anything at all .
(src)=”1.2013”	bonjour , Barbicane .
(src)=”1.2014”	Comment cela va-t-il ?
(trg)=”1.1508”	how d ’ye do , Barbicane ?
(trg)=”1.1509”	how are you getting on ?

Figure 2: Alignment errors. In both cases, two consecutive sentences use similar words, which makes the block alignment look better than a split.

This approach can be improved in many ways: an obvious extension will be to add more features, such as cognates, Part-of-Speech, lemmas, or alignment features as was done in (Munteanu and Marcu, 2005). We plan to provide a much tighter integration with Moore’s algorithm, which already computes such alignments, so as to avoid having to recompute them. Finally, the greedy approach to link selection can easily be replaced with an exact search based on dynamic programming techniques, including dependencies with the left and right alignment links.

Acknowledgments

This work has been partly funded through the “Google Digital Humanities Award” program.

7. References

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Coling 2010: Posters*, pages 81–89, Beijing, China. Coling 2010 Organizing Committee.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991, Berkeley, California*, pages 169–176.

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.

Stanley F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL ’93, pages 9–16.

Yun-Chuang Chiao, Olivier Kraif, Dominique Laurent, Thi Minh Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis, and Wajdi Zaghouani. 2006. Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of the 5th international Conference on Language Resources and Evaluation - LREC’06*, Genoa, Italy.

Yonggang Deng, Shankar Kumar, and William Byrne. 2007. Segmentation and alignment of parallel text for statistical machine translation. *Natural Language Engineering*, 13(03):235–260.

Mohamed Abdel Fattah, David B. Bracewell, Fuji Ren, and Shingo Kuroiwa. 2007. Sentence alignment using P-NNT and GMM. *Computer Speech and Language*, 21:594–608, October.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 177–184, Berkeley, California.

Martin Kay and Martin Röscheisen. 1993. Text-translation alignement. *Computational Linguistics*, 19(1):121–142.

Philippe Langlais. 1998. A System to Align Complex Bilingual Corpora. Technical report, CTT, KTH, Stockholm, Sweden, Sept.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528.

Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *In Proceedings of LREC-2006*, Genoa, Italy.

Elliot Macklovitch. 1994. Using bi-textual alignment for translation validation: the TransCheck system. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 157–168, Columbia.

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

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *Proc. AMTA’02*, Lecture Notes in Computer Science 2499, pages 135–144, Tiburon, CA, USA. Springer Verlag.

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

Ardwait Rathnaparkhi. 1998. *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Ph.D. thesis, University of Pennsylvania.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, November.

Michel Simard and Pierre Plamondon. 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80.

Michel Simard, George F. Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora.

In Ann Gawman, Evelyn Kidd, and Per-Åke Larson, editors, *Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research, October 24-28, 1993, Toronto, Ontario, Canada, 2 Volume*, pages 1071–1082.

Michel Simard. 1998. The BAF: a corpus of English-French bitext. In *First International Conference on Language Resources and Evaluation*, volume 1, pages 489–494, Grenada, Spain.

Jörg Tiedemann. 2011. *Bitext Alignment*. Number 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed). Morgan & Claypool Publishers.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596, Borovets, Bulgaria.

Jean Véronis and Philippe Langlais. 2000. Evaluation of Parallel Text Alignment Systems. In Jean Véronis, editor, *Parallel Text Processing*, Text Speech and Language Technology Series, chapter X, pages 369–388. Kluwer Academic Publishers.

Dekai Wu. 1994. Aligning a parallel English-Chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics, 1994, Las Cruces, New Mexico Canada*, pages 80–87.

Dekai Wu. 2010. Alignment. In Nitin Indurkhya and Fred Damerau, editors, *CRC Handbook of Natural Language Processing*, number 16, pages 367–408. CRC Press.