Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, October 2015

Sentence alignment for literary texts

The state-of-the-art and beyond

Yong Xu
Aurélien Max
François Yvon

Published by CSLI Publications

LiLT volume 12, issue 6 October 2015

Sentence alignment for literary texts

The state-of-the-art and beyond

YONG XU, *LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, yong.xu@limsi.fr*
AURÉLIEN MAX, *LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, amax@limsi.fr*
FRANÇOIS YVON, *LIMSI, CNRS, Université Paris-Saclay, francois.yvon@limsi.fr*

Abstract

Literary works are becoming increasingly available in electronic formats, thus quickly transforming editorial processes and reading habits. In the context of the global enthusiasm for multilingualism, the rapid spread of e-book readers, such as Amazon Kindle® or Kobo Touch®, fosters the development of a new generation of reading tools for bilingual books. In particular, literary works, when available in several languages, offer an attractive perspective for self-development or everyday leisure reading, but also for activities such as language learning, translation or literary studies.

An important issue in the automatic processing of multilingual e-books is the alignment between textual units. Alignment could help identify corresponding text units in different languages, which would be particularly beneficial to bilingual readers and translation professionals. Computing automatic alignments for literary works, however, is a task more challenging than in the case of better behaved corpora such as parliamentary proceedings or technical manuals. In this paper, we revisit the problem of computing high-quality

LiLT Volume 12, Issue 6, October 2015.
Sentence alignment for literary texts.

alignment for literary works. We first perform a large-scale evaluation of automatic alignment for literary texts, which provides a fair assessment of the actual difficulty of this task. We then introduce a two-pass approach, based on a maximum entropy model. Experimental results for novels available in English and French or in English and Spanish demonstrate the effectiveness of our method.

1 Introduction

In the digital era, more and more books are becoming available in electronic form. Widely used devices, such as Amazon Kindle® and Kobo Touch®, have made e-books an accepted reading option for the general public. Works of fiction account for a major part of the e-book market.[1] Global economic and cultural exchange also facilitates the dissemination of literature, and many works of fiction nowadays target an international audience. Successful books are pre-sold and translated very rapidly to reach the largest possible readership.[2] Multiple versions of e-books constitute a highly valuable resource for a number of uses, such as language learning (Kraif and Tutin, 2011) or translation studies.

While reading a novel in the original language often helps better appreciate its content and spirit, a non-native reader may come across many obstacles. Some expressions can be difficult to render or even translate in a foreign language (consider idioms or jargon); some sentences use very complex structures; some paragraphs convey highly subjective and delicate arguments; even worse, some authors tend to use a very rich vocabulary, which is difficult to match for foreign readers. In such circumstances, an alignment between two versions of the same book could prove very helpful (Pillias and Cubaud, to appear, 2015): A reader, upon encountering a difficult fragment in the original language, would be able to refer to its translation in a more familiar language.[3]

In this study, we focus on the task of *sentence alignment* for works of fiction, leaving the study of word alignment for later; see (Wu, 2010, Tiedemann, 2011) for two recent surveys. Sentence alignment aims at identifying correspondence between small groups of sentences in bitexts made of a source text and its corresponding translation: such links often match one sentence with one sentence (henceforth 1:1 links), but more complex alignment is also relatively common. This task, generally considered as relatively easy, has received much attention in the early days of word-based Statistical Machine Translation (SMT), driven by the need to obtain large amounts of parallel sentences to train translation models. Most approaches to sentence alignment are unsupervised and share two important assumptions which help make the problem computationally tractable: (a) a restricted number of *link types* suffice to capture most alignment patterns, the most common types being 1:1 links, then 1:2 or 2:1; (b) the relative order of sentences is the same on the two sides of the bitext. From a bird's eye view, alignment techniques can

[1] About 70 % of the top 50,000 bestselling e-books on Amazon are in the 'fiction' category (source: http://authorearnings.com/report/the-50k-report/).

[2] For instance, J. K. Rowling's *Harry Potter* has already been translated into over 70 languages.

[3] An example implementation is at http://www.doppeltext.com/.

be grouped into two main families: on the one hand, *length-based approaches* (Gale and Church, 1991, Brown et al., 1991) exploit the fact that a short sentence has a short translation, and a long sentence has a long translation. On the other hand, *lexical matching approaches* (Kay and Röscheisen, 1993, Chen, 1993, Simard et al., 1993, Melamed, 1999, Ma, 2006) identify sure anchor points for the alignment using bilingual dictionaries or crude surface similarities between word forms. Length-based approaches are fast but error-prone, while lexical matching approaches seem to deliver more reliable results but at higher computational cost. The majority of the state-of-the-art approaches to the problem (Langlais, 1998, Simard and Plamondon, 1998, Moore, 2002, Varga et al., 2005, Braune and Fraser, 2010, Lamraoui and Langlais, 2013) combine both types of information.

The goal of these methods, however, is primarily to deliver *high-precision* parallel sentence pairs to fuel SMT systems or feed translation memories in specialized domains. They can then safely prune blocks of sentence pairs whenever their alignment is uncertain; some methods even only target high-confidence, 1:1, alignment links. A significant part of recent developments in sentence alignment have tried to make the large-scale harvesting of parallel sentence pairs work also for noisy parallel data (Éva Mújdricza-Maydt et al., 2013), as well as for comparable bilingual corpora (Munteanu and Marcu, 2005, Smith et al., 2010), using *supervised learning techniques*.

While these restrictions are reasonable for the purpose of training SMT systems,[4] for other applications, such as bitext visualization, translator training, automatic translation checking, the alignment for the entire bitext should be computed. This is especially the case for multilingual works of fiction: the parts that are more difficult for automatic alignment algorithms (usually involving highly non-literal translations or large blocks of insertions/deletions) often correspond to parts where a reader might also look for help. For such tasks, it seems that both precision and recall are to be maximized.

This paper therefore reconsiders the task of full-text sentence alignment with two goals: (a) re-evaluate the actual performance of state-of-the-art methods for literary texts, both in terms of their precision and recall, using large collections of publicly available novels; (b) develop, analyse and improve an algorithm initially introduced in (Yu et al., 2012a,b), which was shown to outperform a significant sample of sentence alignment tools on a set of manually aligned corpora.

The rest of this paper is organized as follows: we briefly review, in Section 2, several state-of-the-art sentence alignment tools and evaluate their performance, first on two small reference datasets of gold alignments, then on a

[4]The work by Uszkoreit et al. (2010), however, shows that this procedure actually discards a lot of useful data.

much larger set of *approximately* correct alignments. Section 3 presents our two-pass alignment algorithm; one of its distinguishing feature is its use of external resources to improve its decisions. Experiments on two language pairs (English-French and English-Spanish) are presented and discussed in Section 4, before we recap our main findings and discuss further prospects in Section 5.

2 Aligning literary texts: solved or unsolved?

Commenting on the unsatisfactory results achieved by all sentence alignment systems during the Arcade evaluation campaign (Véronis and Langlais, 2000) on the single test book, Jules Verne's *De la terre à la lune*, Langlais et al. (1998) hint that:

> these poor results are linked to the literary nature of the corpus, where translation is freer and more interpretative,

They express a general feeling that literary texts should be more difficult to align than, say, technical documents. However, assessing the real difficulty of the task is in itself challenging, for lack of a large set of books annotated with a reference (gold) alignment. For instance, the recent study of Éva Mújdricza-Maydt et al. (2013) on English-German alignment used only three books for evaluation. In this section, we aim to provide a more precise answer to this question, using a large collection of *partially aligned* books in two language pairs.

2.1 The state-of-the-art

To evaluate state-of-the-art performance, we first need to identify baseline tools, appropriate evaluation metrics and representative test sets.

Baseline tools

Baseline alignments are computed using several open-source sentence alignment packages: GMA (Melamed, 1999),[5] BMA (Moore, 2002),[6] Hunalign (Varga et al., 2005),[7] Gargantua (shortened to Garg in the tables) (Braune and Fraser, 2010),[8] and Yasa (Lamraoui and Langlais, 2013).[9] These tools constitute, we believe, a representative sample of the current state-of-the-art in sentence alignment. Note that we leave aside here approaches inspired by Information Retrieval techniques such as (Bisson and Fluhr, 2000), which models sentence alignment as a cross-language information retrieval task, or

[5] http://nlp.cs.nyu.edu/GMA/
[6] http://research.microsoft.com/en-us/downloads/aafd5dcf-4dcc-49b2-8a22-f7055113e656/
[7] http://mokk.bme.hu/en/resources/hunalign/
[8] http://sourceforge.net/projects/gargantua/
[9] http://rali.iro.umontreal.ca/rali/?q=en/yasa

Sennrich and Volk's (2010) approach, also based on some automatic translation of the "source" text, followed by a monolingual matching step.

GMA, introduced by Melamed (1999), is the oldest approach included in this sample, and yet one of the most effective: assuming "sure" lexical anchor points in the bitext map, obtained e.g. using bilingual dictionaries or cognate-based heuristics, GMA greedily builds a so-called "sentence map" of the bitext, trying to include as many anchors as possible, while also remaining close to the bitext "diagonal". A post-processing step will take sentence boundaries into account to deliver the final sentence alignment. Note that GMA uses no length cues, and also that it has been shown to perform particularly well at spotting large omissions in a bitext (Melamed, 1996).

Moore's (2002) approach implements a two-pass, coarse-to-fine, strategy: a first pass, based on sentence length cues, computes a first alignment according to the principles of length-based approaches (Brown et al., 1991, Gale and Church, 1991). This initial alignment is used to train a simplified version of IBM model 1 (Brown et al., 1993), which provides the alignment system with lexical association scores. These scores are then used to refine the measure of association between sentences. This approach is primarily aimed at delivering high-confidence, 1:1 sentence alignments to be used as training material for data-intensive MT. Sentences that cannot be reliably aligned are discarded from the resulting alignment.

Hunalign is described in (Varga et al., 2005). It also implements a two-pass strategy which resembles Moore's approach. The main difference is that Hunalign also produces many-to-one and one-to-many alignment links, which are needed to ensure that all the input sentences are actually aligned.

Gargantua (Braune and Fraser, 2010) is, similarly to the approach of Deng et al. (2007) and our own approach, an attempt to improve the final steps of Moore's algorithm. The authors propose a two-pass unsupervised approach that works along the following lines: (a) search for an optimal alignment considering only links made of at most one sentence in each language (including null links); (b) heuristically improve this initial solution by merging adjacent links. A key observation in this work is that step (a) can be very fast. Like most work in this vein, this approach requires explicit modelling of null links, and is prone to miss large untranslated portions on one side of the bitext.

The work of Lamraoui and Langlais (2013) also performs multiple passes over the data, but proceeds in the reverse order: predefined lexical associations (from a bilingual dictionary or from so-called *cognates*) are first used to prune the alignment search space to a restricted number of near-diagonal alignments (the diagonal is defined with respect to these sure *anchor* alignment points.). The second pass will then perform dynamic programming search with the additional help of a length-based model. In spite of its simplicity, this computationally lightweight approach is reported to perform

remarkably well Lamraoui and Langlais (2013).

Evaluation metrics

Sentence alignment tools are usually evaluated using standard *recall* (R) and *precision* (P) measures, combined in the *F-measure* (F), with respect to some manually defined gold alignment (Véronis and Langlais, 2000). These measures can be computed at various levels of granularity: at the level of alignment links, of sentences, of words, and of characters. Because gold references only specify alignment links, the other references are automatically derived in the most inclusive way. As a side effect, all metrics but the link-level ones *ignore null alignments*.[10] Our results are therefore based solely on the link-level F-measure, so as to reflect the importance of correctly predicting unaligned sentences in our targeted applicative scenario.

Evaluation corpora

The performance of sentence alignment algorithms is typically evaluated on reference corpora for which a gold alignment is provided. Manual alignment constitutes the most reliable reference, but is quite rare. In this work, we have used two sets of manually aligned literary works: one is an extract of the BAF corpus (Simard, 1998), consisting of one book by Jules Verne, *De la terre à la lune*; the other has been developed for a preliminary study described by Yu et al. (2012a), and is made up of four novels translated from French into English and three from English into French. These two sets are both relatively small, and only contain bitexts in English and French. These corpora were manually aligned, so they may contain links of arbitrary types. These gold references constitute our main source of evidence for comparing the various algorithms used in this study.

For the purpose of a larger-scale evaluation, we have also made use of two much larger, multi-parallel, corpora of publicly available books that are available on the Internet.[11] The corpus `auto en-fr` contains novels in English and French, and the corpus `auto en-es` contains novels in English and Spanish. These corpora are only imperfectly aligned, and are used to provide approximations of the actual alignment quality. Table 1 provides basic statistics for all these evaluation sets.

Note that there is an obvious discrepancy in the BAF corpus between the number of sentences on the two sides of the bitext, which makes the auto-

[10] Assume, for instance, that the reference alignment links the pair of Foreign sentences ($\mathbf{f}_1$, $\mathbf{f}_2$) to the single English sentence $\mathbf{e}$: reference *sentence-level alignments* will contain both ($\mathbf{f}_1, \mathbf{e}$) and ($\mathbf{f}_2, \mathbf{e}$); likewise, reference *word-level alignments* will contain all the possible word alignments between tokens in the source and the target side, etc. For such metrics, missing the alignment of a large "block" of sentences is more harmful than missing a small one; likewise, misaligning short sentences is less penalized than misaligning longer ones.

[11] See http://www.farkastranslations.com/bilingual_books.php

	# books	lang.	# links	# sent. en	# sent. fr or es
BAF	1	en-fr	2,520	2,554	3,319
`manual en-fr`	7	en-fr	1,790	1,970	2,100
`auto en-fr`	24	en-fr	75,731	129,022	126,561
`auto en-es`	17	en-es	61,181	102,545	104,216

TABLE 1 Corpus statistics.

matic alignment of this book especially challenging. Also note that in the larger corpora, `auto en-fr` and `auto en-es`, manual alignment links are defined at the level of *paragraphs*, rather than at the level of sentences.

Baseline evaluation

We evaluate the performance of the baseline sentence alignment tools on these four corpora, using the standard link-level F-measure. As explained above, the two larger corpora are aligned *at the paragraph level*, meaning that such resources cannot be readily used to compute alignment quality scores. Our solution has been to refine this coarse alignment by running the Gale and Church (1991) alignment program to compute within-paragraph sentence alignments, keeping the paragraph alignments unchanged from the reference. This approach is similar to the procedure used to align the Europarl corpus at the sentence level (Koehn, 2005), where reliable paragraph boundaries are readily derived from speaker turns or session changes. As a result, these semi-automatic references only contain a restricted number of link types as computed by Gale and Church (1991) program: 1:0, 0:1, 1:1, 1:2, and 2:1. We then take these partially correct alignments as pseudo-references for the purpose of evaluating alignment tools – keeping in mind that the corresponding results will only be approximate. Our main evaluation results are in Table 2. For more details, see the Appendix, Tables I, II and III.

Regarding the gold corpus (`manual en-fr`), the numbers in the top part of Table 2 show that the alignment problem is far from solved, with an average F-score around 80 for the three best systems (Gargantua, GMA and Yasa).

On the two larger corpora, a legitimate question concerns the reliability of numbers computed using semi-automatic references. To this end, we manually aligned excerpts of three more books: Lewis Carroll's *Alice's Adventures in Wonderland*, Arthur Conan Doyle's *The Hound of the Baskervilles*, and Edgar Allan Poe's *The Fall of the House of Usher*, for a total of 1,965 sentences on the English side. We then computed the difference in performance observed when replacing the gold alignments with semi-automatic ones. For these three books, the average difference between the evaluation on pseudo-references and on actual references is less than 2 points in F-measure; furthermore, these differences are consistent across algorithms.

		GMA	BMA	Hun	Garg	Yasa
BAF		61.4	73.6	71.2	65.6	75.7
manual en-fr	min	53.5	57.4	54.3	51.7	59.9
	max	92.8	91.5	92.6	97.1	95.6
	mean	79.6	74.9	74.5	80.2	79.1
auto en-fr	min	62.1	47.1	56.6	56.4	62.3
	max	99.5	98.4	99.5	98.1	98.8
	mean	88.7	84.0	87.9	88.7	89.6
auto en-es	min	60.3	48.8	43.7	60.9	58.3
	max	96.5	98	96.4	98.8	98.4
	mean	82.8	78.4	81.0	80.5	82.7

TABLE 2 Baseline evaluation results.

A second comforting observation is that the same ranking of baseline tools is observed across the board: Gargantua, GMA and Yasa tend to produce comparable alignments, outperforming Hunalign and BMA by approximately 2 to 3 F-measure points. It is also worth pointing out that on the two large datasets less than 3% of the sentence links computed by BMA actually cross the reference paragraph boundaries; this warrants our assumption that BMA actually computes sure 1:1 links. Note that even for "easy" books, the performance falls short of what is typically observed for technical documents, with F-measure hardly reaching 0.95; for difficult ones (such as Jane Austen's *Pride and Prejudice*), the best F-measure can be as low as 0.62. From this large-scale experiment, we can conclude that sentence alignment for literary texts remains challenging, even for relatively easy language pairs such as English-French or English-Spanish. It is expected that sentence alignment can only be more difficult when involving languages that are historically or typologically unrelated, or that use different scripts.

3 A Maxent-Based Algorithm

We present in this section our approach to obtaining high-quality alignments. We borrow from Yu et al. (2012a) the idea of a two-pass alignment process: the first pass computes high-confidence $1:1$ links and outputs a partially aligned bitext containing sure links and residual *gaps*, i.e. parallel blocks of non-aligned sentences. These small blocks are then searched using a more computationally costly,[12] but also more precise, model, so as to recover the missing links. We propose, again following our previous work, to evaluate possible intra-block alignments using a maximum entropy (MaxEnt) model – trained here on a large external corpus using a methodology and features sim-

[12] Too costly, in fact, to be used in the first pass over the full bitext.

ilar to (Munteanu and Marcu, 2005, Smith et al., 2010, Éva Mújdricza-Maydt et al., 2013). These steps are discussed below in detail.

3.1 A MaxEnt Model for Parallel Sentences

Any sentence alignment method needs, at some point, to assess the level of parallelism of a sentence pair, based on a surface description of these sentences. As discussed above, two kinds of clues are widely employed in existing systems to perform such assessment: sentence lengths and lexical information. Most dynamic programming-based approaches further impose a prior probability distribution on link types (Gale and Church, 1993).

Our system combines all the available clues in a principled, rather than heuristic, way, using a MaxEnt model.[13] For any pair of sentences $\mathbf{l} = (\mathbf{e}, \mathbf{f})$, the model computes a link posterior probability $p(Y = y|\mathbf{e}, \mathbf{f})$, where Y is a binary variable for the existence of an alignment link. The rationale for using MaxEnt is (a) that it is possible to efficiently integrate as many features as desired into the model, and (b) that we expect the resulting posterior probabilities to be less peaked towards extreme values than what we have observed with generative alignment models such as Moore's model. We give in Section 4.2 the details of the features used in our model.

A second major difference with the existing approaches is our use of a very large set of high-confidence alignment links to train our model. Indeed, most sentence alignment systems are *endogenous*: they only rely on information extracted from the bitext under consideration. While this design choice was probably legitimate in the early 1990s, it is much more difficult to justify now, given the wide availability of sentence-aligned parallel data, such as the Europarl corpus (Koehn, 2005), which can help improve alignment systems. This is in departure from our own past work, where the training data for MaxEnt was identified during the first pass, resulting in small and noisy datasets: Yu et al. (2012a) observe that for an extreme case of poor first-pass alignment, MaxEnt is trained using only four positive examples.

To better match our main focus, which is the processing of literary texts, we collected positive instances from the same publicly available source, extracting all one-sentence paragraphs as reliable alignment pairs. This resulted in a gold set of approximately 18,000 sentences pairs for French/English (out of a grand total of 125,000 sentence pairs). Negative examples are more difficult to obtain and are generated artificially, as explained in Section 4.3.

Finally, note that our model, even though it is trained on 1:1 sentence pairs, can in fact evaluate any pairs of segments. We make use of this property in

[13] We used the implementation from http://homepages.inf.ed.ac.uk/lzhang10/maxent toolkit.html.

our implementation.[14]

3.2 Computing sure 1-1 links

As in many existing tools that implement a multi-step strategy, the main purpose of the first step is to provide a coarse alignment, in order to restrict the search space of the subsequent steps. In our approach, the links computed in the first step are mainly used as anchor points, which makes the more costly search procedure used in the second step feasible. Since we do not reevaluate these anchor links, they should be as reliable as possible.

Our current implementation uses Moore's (2002) algorithm to identify these sure anchors: as explained in Section 2.1, this algorithm tends to obtain a very good precision, at the expense of a less satisfactory recall. Moore's algorithm also computes posterior probabilities for every possible link, which are then used as confidence scores. This system has good precision because it discards all links with a posterior probability lower than 0.5. As explained below, such confidence measures can be used to control the quality of the anchor points used downstream. Table 3 illustrates the result of this first step.

en_1	Poor Alice!	Pauvre Alice!	fr_1
en_2	It was as much as she could do, lying down on one side, to look through into the garden with one eye; but to get through was more hopeless than ever: she sat down and began to cry again.	C'est tout ce qu'elle put faire, après s'être étendue de tout son long sur le côté, que de regarder du coin de l'oeil dans le jardin. Quant à traverser le passage, il n'y fallait plus songer. Elle s'assit donc, et se remit à pleurer.	fr_2 fr_3 fr_4
en_3	"You ought to be ashamed of yourself," said Alice, "a great girl like you," (she might well say this), "to go on crying in this way!	«Quelle honte!» dit Alice. «Une grande fille comme vous» («grande» était bien le mot) «pleurer de la sorte!	fr_5 fr_6
en_4	Stop this moment, I tell you!"	Allons, finissez, vous dis-je!»	fr_7
en_5	But she went on all the same, shedding gallons of tears, until there was a large pool all round her, about four inches deep and reaching half down the hall.	Mais elle continue de pleurer, versant des torrents de larmes, si bien qu'elle se vit à la fin entourée d'une grande mare, profonde d'environ quatre pouces et s'étendant jusqu'au milieu de la salle.	fr_8

TABLE 3 An example alignment computed by Moore's algorithm for *Alice's Adventures in Wonderland*. The first and third anchor links delineate a 2×5 gap containing 2 English and 5 French sentences.

[14] Training the model on 1:1 links, and using it to assess multi-sentence links creates a small methodological bias. Our attempts to include other types of links during training showed insignificant variance in the performance.

3.3 Closing alignment gaps

The job of the second step is to complete the alignment by filling in first-pass gaps. Assume that a gap begins at index i and ends at index j ($i \leq j$) on the English side, and begins at index k and ends at index l ($k \leq l$) on the Foreign side. This step aims at refining the alignment of sentences $\mathbf{e}_{i,j}$ and $\mathbf{f}_{k,l}$, assuming that these blocks are already (correctly) aligned as a whole.

If one side of the gap is empty,[15] then nothing is to be done, and the block is left as is. In all other cases, the block alignment will be improved by finding a set of n links $\{\mathbf{l}_1, \ldots, \mathbf{l}_n\}$ that maximize the following score:

$$\prod_{i=1}^{n} \frac{p(1|\mathbf{l}_i = (\mathbf{e}_i, \mathbf{f}_i))}{\alpha \times size(\mathbf{l}_i)} \quad (1)$$

where $size(\mathbf{l})$ is the size of link $\mathbf{l}$, defined as the product of the number of sentences on the source and target sides, and α is a hyper-parameter of the model. Note that this score is slightly different from the proposal in (Yu et al., 2012a), where we used an additive rather than multiplicative penalty in Equation 1. The score computes the probability of an alignment as the product of the probabilities of individual links. The size-based penalty is intended to prevent the model from preferring large blocks over small ones: this is because the scores of alignments made of large blocks contain fewer factors in Equation 1; dividing by $\alpha \times size(\mathbf{l}_i)$ mitigates this effect and makes scores more comparable.

In general, the number of sentences in a gap makes it possible to consider all the sub-blocks within that gap. In some rare cases, however, there were too many sub-blocks to enumerate, and we had to impose an additional limitation on the number of sentences on both sides of a link. Our inspection of several manually annotated books revealed that links are seldom composed of more than four sentences on either side. We have used this limit in our experiments – this is parameter δ in the algorithms below. Note that this is consistent with previous works such as (Gale and Church, 1993, Moore, 2002) where one only considers small links; the largest alignment links are 2:1. Our two-pass approach allows us to explore more alignment types; the largest link type is 4:4. We compare below two algorithms for finding the optimal set of links: a greedy search presented by Yu et al. (2012b), and a novel (exact) algorithm based on dynamic programming.

Greedy search

Greedy search is described in Algorithm 1. It simply processes the possible links in the decreasing probability order and insert them into the final align-

[15] This happens when the algorithm detects two consecutive anchors (i, k) and $(j, k+1)$ where $j > i+1$ (and similarly in the other direction).

ment unless they overlap with an existing alignment link. For a gap with M English sentences and N Foreign sentences and fixed δ, the worst-case complexity of the search is $O(M \times N)$. Given the typical small size of the gaps (see Figure 1), such search can be computed very quickly.

Algorithm 1 Greedy search

Input: block = $(\mathbf{e}_{i,j}, \mathbf{f}_{k,l})$, priority list L, max gap size δ
Output: result list R
  Generate the set of all possible links S between sub-blocks in $(\mathbf{e}_{i,j}, \mathbf{f}_{k,l})$
  for all l in S **do**
    insert **l** into L, with $score(\mathbf{l})$ defined by (1)
  end for
  while L not empty **do**
    pop top link $\mathbf{l}^*$ from L
    insert $\mathbf{l}^*$ into R
    remove from L any link that intersects or crosses $\mathbf{l}^*$
  end while
  complete R with null links

The result of Algorithm 1 is a collection of links which do not overlap with each other and respect the monotonicity constraint. Note that even though the original list L does not contain any null link, the resulting alignment R may contain links with an empty source or target side. For instance, if links $(\mathbf{e}_{b,b+1}, \mathbf{f}_{d-1,d})$ and $(\mathbf{e}_{b+2,b+3}, \mathbf{f}_{d+2,d+3})$ are selected, then the null link $(, \mathbf{f}_{d+1,d+1})$ will also be added to R.

Dynamic programming search

The other search algorithm considered in this study is based on dynamic programming (DP). Given a series of English sentences $\mathbf{e}_{i,j}$ and the corresponding Foreign sentences $\mathbf{f}_{k,l}$, DP tries to find the set of links that yield maximal global score. Our DP search procedure is described in Algorithm 2. As it is typical of DP approaches, the algorithm merely amounts to filling a table D containing the score of the best alignments of sub-blocks of increasing size. The search complexity for a gap containing M English and N Foreign sentences is $O(M \times N)$. The constant term in the complexity analysis depends on the types of links DP has to consider. As explained above, we only consider here links with fewer than four sentences on each side.

An important issue for DP search is that the probability of null links must be estimated. This is difficult for MaxEnt, because no such information can be found in the training corpus. In greedy search, which only considers non-null links, this problem does not exist. In DP, however, null links appear in all backtraces. We have adopted here a simple method, which is to estimate the

Algorithm 2 The dynamic programming search

Input: Gap=($\mathbf{e}_{i,j}, \mathbf{f}_{k,l}$), empty tables D and B, max gap size δ
Output: link list R

```
for a ← i to j do
  for b ← k to l do
    max ← 0
    for m ← 0 to min(a, δ) do
      for n ← 0 to min(b, δ) do
        cur ← D(a − m, b − n) + score(e_{a−m,a}, f_{b−n,b})
        if cur ≥ max then
          max ← cur
          D(a, b) ← max
          B(a, b) ← (a − m, b − n)
        end if
      end for
    end for
  end for
end for
Back trace on B to find R
```

score of a null link $\mathbf{l} = (, \mathbf{e})$ or $\mathbf{l} = (\mathbf{e},)$ as:

$$score(\mathbf{e},) = score(, \mathbf{e}) = exp(-\beta|\mathbf{e}|) \quad (2)$$

where $|u|$ returns the number of tokens in u and $\beta > 0$ is a hyper-parameter of the method. The intuition is that long null links should be less probable than shorter ones.

4 Experiments

We conducted our experiments on the same corpora as in Section 2. We first discuss the effect of various hyper-parameters of the system on its overall performance. Next, we report the results of our alignment algorithm on these corpora, compared to the results of other methods.

4.1 A study of Moore's alignments (BMA)

Our method relies heavily on the information computed by Moore's algorithm in the first step, since we use those links as anchor points to prune the search space of the second step. The number of anchor points has an effect on the computational burden of the search algorithm. Their quality is even more important because incorrect anchors hurt the performance in two ways. They count as errors in the final result, and they propagate erroneous block alignments for the second step, thereby generating additional alignment errors. It is then natural to investigate the quality of BMA's results from several

perspectives.

On the BAF corpus, which contains a complete reference alignment, Moore's algorithm returns 1944 1:1 links, among which only 1577 are correct (P=0.81). The 1944 links define 445 gaps to be aligned by the second alignment pass. The quality of these gaps is also relevant. We define a gap as correct if it can be fully decomposed into links that appear in the reference set. Among the 445 gaps to be aligned, 180 are incorrect. Finally note that the noise in each incorrect gap also negatively affects the search.

Moore's algorithm associates a confidence score with each link. As shown in Figure 1, using a tighter threshold to select the anchor links significantly improves the precision, and also reduces the number of wrong gaps, at the expense of creating larger blocks.

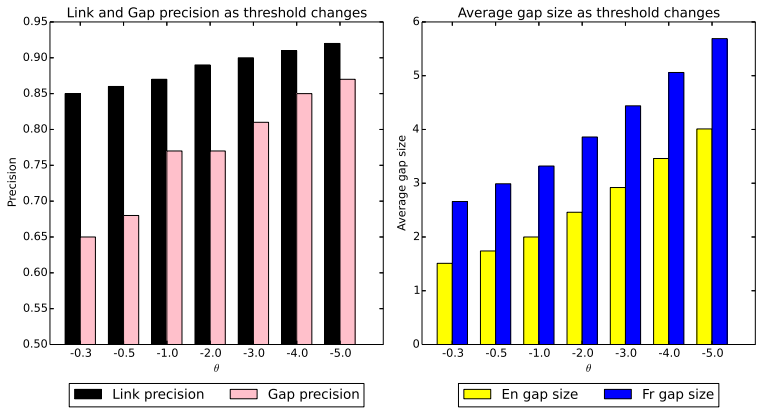

FIGURE 1 Varying confidence threshold causes link precision, gap precision (left) and gap size (right) to change. The numbers are computed over all the manually aligned data (BAF and `manual en-fr`).

In Figure 1, we plot precision as a function of θ, where the actual threshold is equal to 1 - 10^{θ}. For instance, θ = -0.3 corresponds to a threshold of 0.5, and θ = -5 corresponds to a threshold of 0.99999.[16] On BAF, using a very high confidence threshold of 0.99999 improves the precision of anchor points from 0.81 to 0.89, and the ratio of correct gaps rises from 0.6 to 0.82. On the `manual en-fr` corpus of (Yu et al., 2012a), threshold 0.99999 yields an anchor point precision of 0.96 and a correct gap ratio of 0.94. In both cases, a high-confidence threshold significantly reduces the number of wrong gaps. In our implementation, we set the threshold to 0.9999 to reach an acceptable

[16]Posterior link probabilities computed by generative models such as those used by Moore's algorithm tend to be very peaked, which explains the large number of very confident links.

trade-off between correct gap ratio and gap sizes.

4.2 Feature engineering

The core component of our system is a MaxEnt classifier, which, given a pair $(\mathbf{e}, \mathbf{f})$ of source and target sentences, evaluates the level of parallelism between them. In (Yu et al., 2012a), we used a restricted (endogenous) feature set derived solely from the bitext under study. By allowing ourselves also to consider external resources, we can design more complex and effective feature families. This work considers 9 families of features.

1. The *length* in characters of $\mathbf{e}$ and $\mathbf{f}$, and the length ratio, discretized into 10 intervals. This family contains a total of 12 features.
2. The *number of identical tokens* in $\mathbf{e}$ and $\mathbf{f}$. We define 5 features for values (0, 1, 2, 3, 4+).
3. The *number of cognates*[17] in $\mathbf{e}$ and $\mathbf{f}$, also defining 5 features for values (0, 1, 2, 3, 4+).
4. The *word-pair lexical features*, one for each pair of words co-occurring at least once in a parallel sentence. For example, if the first token in $\mathbf{e}$ is "Monday" and the first token in $\mathbf{f}$ is "Lundi", then the pair "Monday-Lundi" defines a word-pair lexical feature.
5. *Sentence translation* score features. For a pair of sentence $\mathbf{e} = \mathbf{e}_1^I$ and $f = \mathbf{f}_1^J$, we use the IBM Model 1 score (Brown et al., 1993):

$$T_1(\mathbf{e}, \mathbf{f}) = \frac{1}{J}\sum_{j=1}^{J}\log(\frac{1}{I} * \sum_{i=1}^{I} p(f_j|e_i))$$

$$T_2(\mathbf{e}, \mathbf{f}) = \frac{1}{I}\sum_{i=1}^{I}\log(\frac{1}{J} * \sum_{j=1}^{J} p(e_i|f_j))$$

   After discretizing these values, we obtain 10 features for each direction.

6. *Longest continuous covered span* features. A word $\mathbf{e}_i$ is said to be covered if there exists one word $\mathbf{f}_j$ such that the translation probability $t(\mathbf{e}_i|\mathbf{f}_j)$ in the IBM Model 1 table is larger than a threshold (10^{-6} in our experiments). A long span of covered words is an indicator of parallelism. We compute the length of the longest covered spans on both sides, and normalize them by their respective sentence lengths. This family contains 20 features.
7. *Uncovered words*. The notion of coverage is defined as above. We count the number of uncovered words on both sides and normalize by the sentence length. This family contains 20 features, 10 on each side.

[17] We call a pair of words "cognates" if they share a prefix of at least 4 characters.

8. *Unlinked words* in the IBM 1 alignment. A word $\mathbf{e}_i$ is said to be linked if in an alignment $\mathbf{a}$ there exists some index j such that $\mathbf{a}_j = i$.[18] Large portions of consecutive unlinked words is a sign of non-parallelism. These counts are normalized by the sentence length, and yield 20 additional features.

9. *Fertility* features. The fertility of a word $\mathbf{e}_i$ is the number of indices j that satisfies $\mathbf{a}_j = i$. Large fertility values indicate non-parallelism. We take, on each side, the three largest fertility values, and normalize them with respect to the sentence lengths. This yields 60 supplementary features.

The feature families 1-4 are borrowed from (Yu et al., 2012a), and are used in several other studies on supervised sentence alignment, e.g. (Munteanu and Marcu, 2005, Smith et al., 2010). All other features rely on IBM Model 1 scores and can only be computed reliably on large (external) sources of data. To evaluate the usefulness of these various features, we performed an incremental feature selection procedure. We first trained the model with only one feature family, then added the other families one by one, monitoring the performance of the MaxEnt model as more features are included. For this study, model performance was measured by the prediction accuracy: the ratio of examples (positive and negative) for which the model makes the right classification. Because the new features are all based on IBM Model 1, the size of the training corpus also has an important effect.

We have thus set up three datasets of increasing sizes. The first set contains around 110,000 tokens, which is the typical amount of data that Moore's algorithm would return upon aligning a single book; the second one contains 1,000,000 tokens; the third one includes all the parallel literary data collected for this study and totals more than 5,000,000 tokens. Each data set is split into a training set, a development set and a test set, using 80% for training, 10% for tuning, and 10% for testing. For these experiments, the model is trained with 30 iterations of the L-BFGS algorithm with a Gaussian prior, which is tuned on the development set.

Table 4 gives the performance of the MaxEnt model on the test set as more features are included. Note that families 2 and 3 are added together.

As expected, the new families of features (5-9) do not help much when trained on a small data set; as more training data are included in the model, the accuracy increases, allowing the system to more than halve the error rate in comparison to the best small data condition.

[18] This notion is different from coverage and assumes that an optimal $1:1$ word alignment has been computed based on IBM 1 model scores. Words can be covered, yet unlinked, when all their possible matches are linked to other words.

	Model accuracy		
	~110K tokens	~1M tokens	~5M tokens
Family 1	0.778	0.873	0.859
+Family 2 and 3	0.888	0.869	0.879
+Family 4	0.957	0.976	0.977
+Family 5	0.943	0.985	0.987
+Family 6	0.912	0.979	0.986
+Family 7	0.913	0.975	0.986
+Family 8	0.913	0.979	0.988
+Family 9	0.913	0.981	0.988

TABLE 4 Evaluation of MaxEnt with varying feature families and training data.

In our applicative scenario, not only do we want the model to make the right alignment decisions, but also expect that it can do so with a high confidence. To check that this is actually the case, we plot the ROC (Receiver Operating Characteristic) curve in Figure 2. We only display the ROC curves for the medium and large data sets.

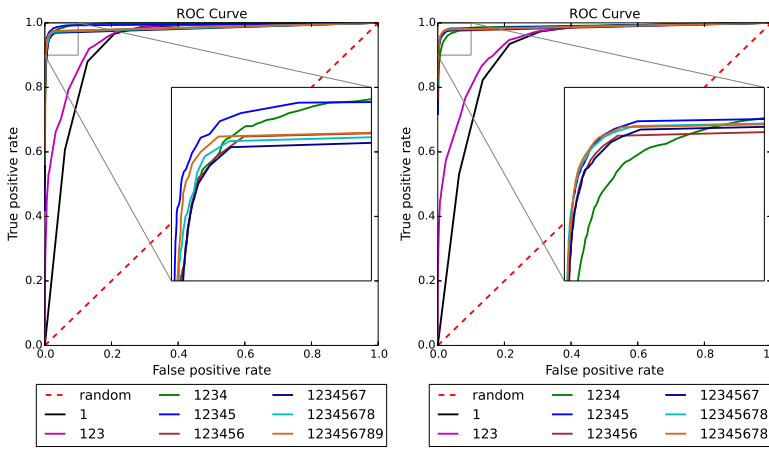

FIGURE 2 ROC curves on the medium size (left) and large (right) data sets. The embedded box is a zoom over the top-left corner.

From the two ROC curves, we can observe that on the medium-size data set, the model achieves very good performance (large AUC areas) in all settings. In both figures, we can see that the use of feature families 6-9 barely improve the confidence of the model over the results of the first five. In the expriments reported below, we only use the feature families 1-5.

4.3 The final system

The final system is constructed as follows (the procedure is identical in both languages pairs). Each test document is processed independently from the others, in a leave-one-out fashion. We use 80% of the sure parallel sentence pairs from the other 23 books for the `auto en-fr` corpus as positive examples, and 16 books for `auto en-es`; we also include the 1:1 links computed by BMA in the left-out corpus. For every pair of parallel sentences $\mathbf{f}_i$ and $\mathbf{e}_j$, we randomly sample three target (respectively, source) sentences other than $\mathbf{e}_j$ (respectively, $\mathbf{f}_i$) to construct a negative pair with $\mathbf{f}_i$ (respectively, $\mathbf{e}_j$). Each positive example thus gives rise to six negative ones. We train a distinct MaxEnt model for each test book, using 80% of these book-specific data, again using 20% of examples as a held-out data set. Only the feature families 1-5 are included in the model, yielding an average test accuracy of 0.988. The ME+DP and ME+gr procedures are then finally used to complete BMA's links according respectively to algorithms 2 and 1. The results are summarized in Table 5 (see also Tables II and III in the Appendix for a full listing). For corpora containing multiple books, the average, minimum and maximum scores are reported.

		GMA	Hun	Garg	Yasa	ME+gr	ME+DP
BAF		61.4	71.2	65.6	75.7	76.3	66.5
`manual en-fr`	min	53.5	54.3	51.7	59.9	61.4	51.0
	max	92.8	92.6	97.1	95.6	95.3	98.0
	mean	79.6	74.5	80.2	79.1	78.3	**81.5**
`auto en-fr`	min	62.1	56.6	56.4	62.3	56.7	57.7
	max	99.5	99.5	98.1	98.8	97.5	97.9
	mean	88.7	87.9	88.7	89.6	85.7	88.9
`auto en-es`	min	60.3	43.7	58.4	64.3	60.4	65.2
	max	96.5	96.4	96.8	100	97.7	98.0
	mean	82.8	81.0	82.6	84.6	80.5	82.7

TABLE 5 Performance of the MaxEnt approach with greedy search (ME+gr) and dynamic programming search (ME+DP) and of four baseline alignment tools.

Our system obtains the best overall results for the manually aligned `manual en-fr` corpus. All the average differences between ME+DP and the other algorithms are significant at the 0.01 level, except for Gargantua and GMA, where the difference is only significant at the 0.05 level. On the large approximate reference sets, ME+DP achieves results comparable with Gargantua and GMA, slightly worse than Yasa. Comparing greedy search with DP search, the mean performance on the `manual en-fr` has been increased by 3 points in F-measure and the differences on the two large corpora,

even though they are only approximations, are also important. Surprisingly, this does not reflect on BAF, where the heuristic search is better than the DP algorithm – again, this might be because of the peculiar trait of BAF, which contains on average larger gaps than the other books (and crucially has an average gap size greater than 4 in one dimension).

We finally performed an error analysis on the manual alignment data set (`manual en-fr`). Table 6 lists the link types in the reference, along with the numbers of reference links that greedy search or DP fails to find.

Link type	in Ref.	ME+gr	ME+DP
0:1	20	13	18
1:0	21	12	18
1:1	1364	68	105
1:2	179	60	36
1:3	32	17	9
2:1	96	54	32
2:2	24	22	19
others	27	22	15
total	1,763	268	252

TABLE 6 Analyses of the errors of greedy search (ME+gr) and DP search (ME+DP) by link type, relative to the number of reference links (in Ref.), for the `manual en-fr` corpus. Only the link types occurring more than 5 times are reported. This filters out 27 links out of 1790.

The numbers in Table 6 suggest that null links remain difficult, especially for the DP algorithm, reflecting the fact that estimating the scores of these links is a tricky issue. This problem arises for all systems whose search is based on DP. For instance, Yasa makes a comparable number of errors for null links (16 errors for type $0:1$, 17 for type $1:0$), Hunalign's results are worse (20 errors for type $0:1$, 19 for type $1:0$), while Gargantua does not return any null link at all. DP tends to be more precise for larger blocks such as 1:2 or 2:1. Table 7 illustrates this property of DP search: this excerpt from Jean-Jacques Rousseau's *Les Confessions* is difficult because of the presence of consecutive 1-to-many links. ME+DP is the only algorithm which correctly aligns the full passage.

The gap size also has an effect on the performance of DP. In DP-search, we constrain alignment links to contain at most 4 sentences on each side, if at least one side of an actual alignment link exceeds this limit. So, our algorithm will fail to find the correct solution. Table 8 displays the average gap size[19]

[19] An $N \times M$ gap contains N sentences on the source side and M sentences on the target side.

en_1	My mother had a defence more powerful even than her virtue; she tenderly loved my father, and conjured him to return; his inclination seconding his request, he gave up every prospect of emolument, and hastened to Geneva.	Ma mère avait plus que la vertu pour s'en défendre; elle aimait tendrement son mari. Elle le pressa de revenir: il quitta tout, et revint.	fr_1 fr_2
en_2	I was the unfortunate fruit of this return, being born ten months after, in a very weakly and infirm state; my birth cost my mother her life, and was the first of my misfortunes.	Je fus le triste fruit de ce retour. Dix mois après, je naquis infirme et malade. Je coûtai la vie à ma mère, et ma naissance fut le premier de mes malheurs.	fr_3 fr_4 fr_5
en_3	I am ignorant how my father supported her loss at that time, but I know he was ever after inconsolable.	Je n'ai pas su comment mon père supporta cette perte, mais je sais qu'il ne s'en consola jamais.	fr_6

TABLE 7 A passage of a reference alignment from Jean-Jacques Rousseau's *Les confessions*. MaxEnt with DP finds all three links.

inside each book of the `manual en-fr` corpus, along with the F-score of greedy search and DP search.

	Ave. gap size	ME+gr	ME+DP
Du Côté de chez Swann	2.62×2.54	89.4	93.3
Emma	10.25×6.75	61.4	51.0
Jane Eyre	4×4.8	67.4	78.9
La Faute de l'Abbé Mouret	1.85×2.79	95.3	98.0
Les Confessions	2.89×4.7	67.8	74.0
Les Travailleurs de la Mer	3.37×3.74	80.8	85.3
The Last of the Mohicans	2.14×3.07	85.8	90.1

TABLE 8 Gap size and performance of MaxEnt on `manual en-fr`.

We can see that DP works better when gap sizes are smaller than 4 on each side. When this is not the case, the results tend to decrease significantly, as for instance for Jane Austen's *Emma*. Greedy search, while generally outperformed by DP, is significantly better for this book. This underscores the need also to improve the anchor detection algorithm in our future work, in order to make sure that gaps are both as correct and as small as possible.

5 Conclusion

This paper has presented a large-scale study of sentence alignment using a small corpus of reference alignments, and two large corpora containing dozens of coarsely aligned copyright-free novels for English-Spanish and English-French language pairs. We have shown that these coarse alignments, once refined, were good enough to compute approximate performance measures for the task at hand, and confirmed the general intuition that automatic

sentence alignment for novels was still far from perfect; some translations appeared to be particularly difficult to align with the original text for all existing methods. Borrowing ideas from previous studies on unsupervised and supervised sentence alignment, we have proposed and evaluated a new alignment algorithm, and showed that it performs better than several strong baselines – even if there remains a lot of room for improvement.

We now plan to study additional, arguably more complex, language pairs to get a more complete picture of the actual complexity of sentence alignment. There also are several obvious weaknesses in our current implementation that we intend to fix. First, it seems unnecessary to continue performing the second step of Moore's algorithm (which basically trains endogenously an IBM 1 Model) because the MaxEnt model also requires IBM 1 scores, which are computed on a large set of clean sentence alignments. Second, the MaxEnt model is trained on isolated sentences and tested with blocks containing one or several sentences; it would be more natural to train the model in the same conditions as observed in testing. Third, there are obvious dependencies between consecutive links that could also be taken into account, changing the MaxEnt with a more complex CRF model (Éva Mújdricza-Maydt et al., 2013). Finally, and importantly, our model needs to compute scores for null links, a nearly impossible task since "true" deletions are difficult to predict based only on the text; we therefore plan to reconcile our approach with techniques, which, like GMA (Melamed, 1999), do not need to model null links, so as to make it more resilient to large unaligned parts.

Acknowledgments

This work has been partially supported by the TransRead project (ANR CONTINT 2011) funded by the French National Research Agency. We have made good use of the alignment data from http://farkastranslations.com/ (©2014), so we would like to thank András Farkas for making his multi-parallel corpus of manually aligned books publicly available.

References

Bisson, Frédérique and Christian Fluhr. 2000. Sentence alignment in bilingual corpora based on crosslingual querying. In *Proceedings of RIAO'2000*, pages 529–542. Paris, France.

Braune, Fabienne 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.

Brown, Peter F., 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.

Brown, Peter F., 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.

Chen, Stanley F. 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.

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

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

Gale, William A. and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics* 19(1):75–102.

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

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. *MT summit* 5:79–86.

Kraif, Olivier and Agnès Tutin. 2011. Using a bilingual annotated corpus as a writing aid: An application for academic writing for efl users. In I. N. K. (Ed.), ed., *Corpora, Language, Teaching, and Resources: From Theory to Practice. Selected papers from TaLC7, the 7th Conference of Teaching and Language Corpora*. Bruxelles: Peter Lang.

Lamraoui, Fethi and Philippe Langlais. 2013. Yet Another Fast, Robust and Open Source Sentence Aligner. Time to Reconsider Sentence Alignment? In *Proceedings of the XIV Machine Translation Summit*, pages 77–84. Nice, France.

Langlais, Philippe. 1998. A System to Align Complex Bilingual Corpora. Tech. rep., CTT, KTH, Stockholm, Sweden.

Langlais, Philippe, Michel Simard, and Jean Véronis. 1998. Methods and Practical Issues in Evaluating Alignment Techniques. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 711–717. Montreal, Quebec, Canada.

Ma, Xiaoyi. 2006. Champollion: A Robust Parallel Text Sentence Aligner. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2006)*. Genoa, Italy.

Melamed, I. Dan. 1996. Automatic Detection of Omissions in Translations. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 764–769.

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

Moore, Robert C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In S. D. Richardson, ed., *Proceedings of the annual meeting of the Association for Machine Translation in the Americas (AMTA 2002)*, Lecture Notes in Computer Science 2499, pages 135–144. Tiburon, CA, USA: Springer Verlag.

Munteanu, Dragos Stefan and Daniel Marcu. 2005. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics* 31(4):477–504.

Pillias, Clément and Pierre Cubaud. to appear, 2015. Bilingual Reading Experiences: What They Could Be and How To Design for Them. In *Proceedings of IFIP Interact 2015*. Bamberg, Germany.

Sennrich, Rico 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)*. Denver, CO.

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

Simard, Michel, George F. Foster, and Pierre Isabelle. 1993. Using cognates to align sentences in bilingual corpora. In A. Gawman, E. Kidd, and P.-Å. Larson, eds., *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.

Simard, Michel and Pierre Plamondon. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation* 13(1):59–80.

Smith, Jason R., Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411.

Tiedemann, Jörg. 2011. *Bitext Alignment*. No. 14 in Synthesis Lectures on Human Language Technologies, Graeme Hirst (ed). Morgan & Claypool Publishers.

Uszkoreit, Jakob, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109. Beijing, China.

Varga, Dániel, 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.

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

Wu, Dekai. 2010. Alignment. In N. Indurkhya and F. Damerau, eds., *CRC Handbook of Natural Language Processing*, pages 367–408. CRC Press.

Yu, Qian, Aurélien Max, and François Yvon. 2012a. Aligning Bilingual Literary Works: a Pilot Study. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 36–44. Montréal, Canada.

Yu, Qian, Aurélien Max, and François Yvon. 2012b. Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC)*. Istanbul, Turkey.

Éva Mújdricza-Maydt, Huiqin Körkel-Qu, Stefan Riezler, and Sebastian Padó. 2013. High-Precision Sentence Alignment by Bootstrapping from Wood Standard Annotations. *The Prague Bulletin of Mathematical Linguistics* 99:5–16.

Appendix

Table I displays the performance of baseline tools on the two manually aligned reference corpora BAF and `manual en-fr`. In Table II, we show the performance of baseline tools on the large corpus `auto en-fr`. The results on the corpus `auto en-es` are provided in Table III. In these tables, we denote "Gr" the MaxEnt-based alignment algorithm with greedy search, and "DP" the MaxEnt-based alignment with dynamic programming.

	Link level F-score						
	GMA	BMA	Hun	Garg	Yasa	Gr	DP
BAF	61.4	73.6	71.2	65.6	75.7	76.3	66.5
Du Côté de chez Swann	92.8	91.5	90.9	92.2	92.2	89.4	93.3
Emma	53.5	57.4	57.7	51.7	59.9	61.4	51.0
Jane Eyre	77.1	61.1	59.3	71.8	66.9	67.4	78.9
La Faute de l'Abbé Mouret	91.5	88.4	92.6	97.1	95.6	95.3	98.0
Les Confessions	71.9	59.6	54.3	68.6	66.7	67.8	74.0
Les Travailleurs de la Mer	80.8	83.4	79.9	87.3	83.8	80.8	85.3
The Last of the Mohicans	89.9	82.7	87.1	92.7	88.6	85.8	90.1

TABLE I F-scores on gold references.

	Link level F-score						
	GMA	BMA	Hun	Garg	Yasa	Gr	DP
20000 Lieues sous les Mers	97.6	96.4	97.2	98.1	98.8	95.9	97.7
Alice's Adventures in Wonderland	81.2	74.3	80.0	83.2	82.7	76.2	81.5
A Study in Scarlet	89.0	78.2	83.8	85.2	89.0	85.0	86.4
Candide	85.7	78.8	82.5	82.6	87.9	79.9	86.4
Germinal	97.2	94.7	97.5	97.6	97.3	95.1	97.9
La Chartreuse de Parme	97.1	94.1	96.1	96.8	97.4	94.2	97.0
La Dame aux Camelias	94.0	91.1	89.6	93.8	94.9	90.7	94.6
Le Grand Meaulnes	93.3	91.0	93.4	94.7	94.1	92.6	94.3
Le Rouge et le Noir	96.9	94.7	96.4	97.2	97.9	94.3	97.3
Les Trois Mousquetaires	88.0	83.3	89.2	88.0	89.9	83.6	87.9
Le Tour du Monde En 80 Jours	76.4	63.9	74.9	75.8	78.5	68.9	75.8
L'île Mystérieuse	93.4	93.3	96.0	94.5	94.8	93.5	94.6
Madame Bovary	93.9	90.7	93.9	94.5	94.1	91.8	95.0
Moll Flanders	80.5	76.9	83.1	81.8	83.3	78.0	82.7
Notre Dame de Paris	93.5	91.1	92.8	94.2	94.1	90.6	94.2
Pierre et Jean	91.4	89.3	91.5	91.9	91.5	88.6	90.7
Pride and Prejudice	62.1	47.1	56.6	56.4	62.3	56.7	57.7
Rodney Stone	88.6	83.7	90.2	90.7	90.1	85.5	89.3
The Fall of The House of Usher	99.5	98.4	99.5	97.4	98.4	97.5	95.7
The Great Shadow	81.7	74.9	83.4	84.0	82.8	79.4	86.0
The Hound of The Baskervilles	92.8	90.5	92.5	93.7	93.5	91.1	93.0
Therese Raquin	85.6	80.3	84.7	85.4	85.1	82.0	86.7
Three Men in a Boat	85.3	76.5	81.7	85.6	87.1	81.8	86.2
Voyage au Centre de la Terre	84.9	81.6	83.7	86.8	85.9	83.5	84.0
Mean	88.7	84.0	87.9	88.7	89.6	85.7	88.9

TABLE II F-scores on the large approximate reference set `auto en-fr`.

	Link level F-score						
	GMA	BMA	Hun	Garg	Yasa	GR	DP
20000 Lieues sous les Mers	88.9	85.2	89.6	89.5	90.2	84.9	89.2
Alice's Adventures in Wonderland	74.2	65.3	66.7	72.2	76.2	71.4	74.4
Anna Karenina Volume I	72.9	69.3	75.4	77.8	73.6	70.9	74.2
Anna Karenina Volume II	69.5	68.2	73.3	75.6	73.2	69.3	72.4
A Study in Scarlet	94.3	91.3	94.1	96.3	96.9	93.1	92.8
Candide	76.8	64.2	71.7	63.2	80.4	71.7	72.8
Die Verwandlung	88.6	80.3	84.9	85.6	87.2	83.1	88.8
Don Quijote de La Mancha	87.9	82.1	85.6	88.8	89.6	81.9	86.6
Jane Eyre	60.3	48.8	43.7	58.4	64.3	60.9	58.3
Les Trois Mousquetaires	82.3	79.2	82.9	83.8	83.3	77.6	83.0
Le Tour du Monde en 80 Jours	78.8	71.1	77.8	77.4	81.4	74.4	78.2
L'île Mystérieuse	77.9	82.5	81.2	80.7	80.1	81.1	79.4
Sense and Sensibility	92.4	88.6	89.9	92.0	93.4	87.8	91.6
The Adventures of Sherlock Holmes	93.7	91.8	93.0	91.3	93.8	91.9	93.5
The Fall of the House of Usher	94.6	98.0	96.4	94.9	100.0	98.8	98.4
The Hound of the Baskervilles	96.5	95.3	95.4	96.8	97.6	95.4	96.2
Voyage au Centre de la Terre	77.2	72.2	75.6	79.1	77.5	74.7	76.4
Mean	82.8	78.4	81.0	82.6	84.6	80.5	82.7

TABLE III F-scores on the large approximate reference set `auto en-es`.