Aligning Bilingual Literary Works: a Pilot Study

Qian Yu and **Aurélien Max** and **François Yvon**
LIMSI/CNRS and Univ. Paris Sud
rue John von Neumann F-91 403 Orsay, France
{fistname.lastname}@limsi.fr

Abstract

Electronic versions of literary works abound on the Internet and the rapid dissemination of electronic readers will make electronic books more and more common. It is often the case that literary works exist in more than one language, suggesting that, if properly aligned, they could be turned into useful resources for many practical applications, such as writing and language learning aids, translation studies, or data-based machine translation. To be of any use, these bilingual works need to be aligned as precisely as possible, a notoriously difficult task. In this paper, we revisit the problem of sentence alignment for literary works and explore the performance of a new, multi-pass, approach based on a combination of systems. Experiments conducted on excerpts of ten masterpieces of the French and English literature show that our approach significantly outperforms two open source tools.

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 the input 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 usual to align paragraphs, sentences, phrases or words (see (Wu, 2010; Tiedemann, 2011) for recent reviews). Alignments are used in many fields, ranging from Translation Studies and Computer Assisted Language Learning (CALL) to Multilingual Natural Language Processing (NLP) applications (Cross-Lingual Information Retrieval, Writing Aids for Translators, Multilingual Terminology Extraction and Machine Translation (MT)). For all these applications, sentence alignments have to be computed.

Sentence alignment is generally thought to be fairly easy and many efficient sentence alignment programs are freely available[1]. Such programs rely on two main assumptions: (i) the relative order of sentences is the same on the two sides of the bitext, and (ii) sentence parallelism can be identified using simple surface cues. Hypothesis (i) warrants efficient sentence alignment algorithms based on dynamic programming techniques. Regarding (ii), various surface similarity measures have been proposed: on the one hand, *length-based* measures (Gale and Church, 1991; Brown et al., 1991) rely on 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 state-of-the-art approaches use both types of information (Langlais, 1998; Simard and Plamondon, 1998; Moore, 2002; Varga et al., 2005; Braune and Fraser, 2010).

In most applications, only high-confidence one-to-one sentence alignments are considered useful and kept for subsequent processing stages. Indeed, when the objective is to build subsentential align-

[1] See, for instance, the Uplug toolbox which integrates several sentence alignment tools in a unified framework: http://sourceforge.net/projects/uplug/

Workshop on Computational Linguistics for Literature, pages 36–44,
Montréal, Canada, June 8, 2012.

ments (at the level of words, terms or phrases), 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 of finer-grained alignments. However, for certain types of bitexts[2], such as literary texts, translation often departs from a straight sentence-by-sentence alignment and using such a constraint can discard a significant proportion of the bitext. For MT, this is just a regrettable waste of potentially useful training material (Uszkoreit et al., 2010), all the more so as parallel literary texts constitute a very large reservoir of parallel texts online. For other applications implying to mine, visualize or read the actual translations in their context (second language learning (Nerbonne, 2000; Kraif and Tutin, 2011), translators training, automatic translation checking (Macklovitch, 1994), etc.), the entire bitext has to be aligned. Furthermore, areas where the translation is only partial or approximative need to be identified precisely.

The work reported in this study aims to explore the quality of existing sentence alignment techniques for literary work and to explore the usability of a recently proposed multiple-pass approach, especially designed for recovering many-to-one pairings. In a nutshell, this approach uses 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) and on a small literary corpus consisting of ten books show that this approach produces high quality alignments and also identifies the most problematic passages better than its competitors.

The rest of this paper is organized as follows: we first report the results of a pilot study aimed at aligning our corpus with existing alignment methods (Section 2). In Section 3, we briefly describe our two-pass method, including some recent improvements, and present experimental performance on the BAF corpus. Attempts to apply this technique to our larger literary corpus are reported and discussed in Section 4. We discuss further prospects and conclude in Section 5.

[2] Actual literary bitexts are not so easily found over the Internet, notably due to (i) issues related to variations in the source text and (ii) issues related to the variations, over time, of the very notion of what a translation should be like.

2 Book alignment with off-the-shelf tools

2.1 A small bilingual library

The corpus used in this study contains a random selection of ten books written mostly in the 19^{th} and in the early 20^{th} century: five are English classics translated into French, and five are French classics translated into English. These books and their translation are freely available[3] from sources such as the Gutenberg project[4] or wikisource[5], and are representative of the kinds of collections that can be easily collected from the Internet. These texts have been preprocessed and tokenized using in-house tools, yielding word and sentence counts in Table 1.

2.2 Baseline sentence alignments

2.2.1 Public domain tools

Baseline alignments are computed using two open-source sentence alignment packages, the sentence alignment tool of Moore (2002)[6], and Hunalign (Varga et al., 2005). These two tools were chosen as representative of the current state-of-the-art in sentence alignment. Moore's 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 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, one-to-one, sentence alignments to be used as training material for data-intensive MT. Sentences that cannot be reliably aligned are discarded from the resulting alignment.

[3] Getting access to more recent books (or their translation) is problematic, due to copyright issues: literary works fall in the public domain 70 years after the death of their author.

[4] http://www.gutenberg.org

[5] http://wikisource.org

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

		French side		English side	
		# sents	# words	# sents	# words
English books and their French translation					
Emma, J. Austen	EM	5,764	134,950	7,215	200,223
Jane Eyre, C. Brontë	JE	9,773	240,032	9,441	237,487
The last of the Mohicans, F. Cooper	LM	6,088	189,724	5,629	177,303
Lord Jim, J. Conrad	LJ	7962	175,876	7,685	162,498
Vanity fair, W. Thackeray	VF	14,534	395,702	12,769	372,027
French books and their English translation					
Les confessions, J.J. Rousseau	CO	9,572	324,597	8,308	318,658
5 semaines en ballon, J. Verne	5S	7,250	109,268	7,894	121,231
La faute de l'Abbé Mouret, E. Zola	AM	8,604	156,514	7,481	156,692
Les travailleurs de la mer, V. Hugo	TM	10,331	170,015	9,613	178,427
Du côté de chez Swann, M. Proust	SW	4,853	208,020	4,738	232,514
	Total	84,731	2,104,698	80,773	2,157,060

Table 1: A small bilingual library

Hunalign[7], with default settings, also implements a two-pass strategy which resembles the approach of Moore. Their 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 appear in the final alignment.

Both systems also deliver confidence measures for the automatic alignment: a value between 0 and 1 for Moore's tool, which can be interpreted as a posterior probability; the values delivered by Hunalign are less easily understood, and range from -1 to some small positive real values (greater than 1).

2.2.2 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: the level of alignment links, of sentences, of words, and of characters. As gold references only specify alignment links, the other references are automatically derived in the most inclusive way. For instance, if the reference alignment links state that the pair of source sentences $\mathbf{f}_1$, $\mathbf{f}_2$ is aligned with target $\mathbf{e}$, the reference sentence alignment will contain both $(\mathbf{f}_1, \mathbf{e})$ and $(\mathbf{f}_2, \mathbf{e})$; likewise, the reference word alignment will contain all the possible word alignments between tokens in the source and the target side. For such metrics, missing the alignment of a large "block" of sentences gets a higher penalty than missing a small one; likewise, misaligning short sentences is less penalized than misaligning longer ones. As a side effect, all metrics, but the more severe one, *ignore null alignments*. Our results are therefore based on the link-level and sentence-level F-measure, to reflect the importance of correctly predicting unaligned sentences in our applicative scenario.

2.2.3 Results

Previous comparisons of these alignment tools on standard benchmarks have shown that both typically yield near state-of-the-art performance. For instance, experiments conducted using the literary subpart of the BAF corpus (Simard, 1998), consisting of a hand-checked alignment of the French novel *De la Terre à la Lune* (*From the Earth to the Moon*), by Jules Verne, with a slightly abridged translation available from the Gutenberg project[8], have yielded the results in Table 2 (Moore's system was used with its default parameters, Hunalign with the `--realign` option).

All in all, for this specific corpus, Moore's strategy delivers slightly better sentence alignments than

[7] ftp://ftp.mokk.bme.hu/Hunglish/src/hunalign; we have used the version that ships with Uplug.

[8] http://www.gutenberg.org/ebooks/83

	P	R	F	% 1-1 links
Alignment based metrics				
Hunalign	0.51	0.60	0.55	0.77
Moore	0.85	0.65	0.74	1.00
Sentence based metrics				
Hunalign	0.76	0.70	0.73	-
Moore	0.98	0.62	0.76	-

Table 2: Baseline alignment experiments

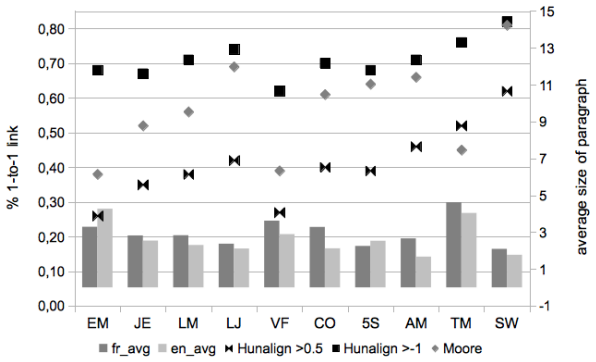

Figure 1: Percentage of one-to-one links and pseudo-paragraph size for various baselines

Hunalign does; in particular, it is able to identify 1-to-1 links with a very high precision.

2.3 Aligning a small library

In a first series of experiments, we simply run the two alignment tools on our small collection to see how much of it can be aligned with a reasonable confidence. The main results are reproduced in Figure 1, where we display both the number of 1-to-1 links extracted by the baselines (as dots on the Figure), as well as the average size of pseudo-paragraphs (see definition below) in French and English. As expected, less 1-to-1 links almost always imply larger blocks.

As expected, these texts turn out to be rather difficult to align: in the best case (*Swann's way* (SW)), only about 80% of the total sentences are aligned by Moore's system; in the more problematic cases (*Emma* (EM) and *Vanity Fair* (VF)), more than 50% of the book content is actually thrown away when one only looks at Moore's alignments. Hunalign's results look more positive, as a significantly larger number of one-to-one correspondences is found. Given that this system is overall less reliable than Moore's approach, it might be safe to filter these alignments and keep only the surer ones (here, keeping only links having a score greater than 0.5). The resulting number of sentences falls way below what is obtained by Moore's approach.

To conclude, both systems seem to have more difficulties with the literary material considered here than with other types of texts. In particular, the proportion of one-to-one links appears to be significantly smaller than what is typically reported for other genres; note, however, that even in the worst case, one-to-one links still account for about 50% of the text. Another finding is that the alignment scores which are output are not very useful: for Moore, filtering low scoring links has very little effect; for Hunalign, there is a sharp transition (around a threshold of 0.5): below this value, filtering has little effect; above this value, filtering is too drastic, as shown on Figure 1.

3 Learning sentence alignments

In this section, we outline the main principles of the approach developed in this study to improve the sentence alignments produced by our baseline tools, with the aim to salvage as many sentences as possible, which implies to come up with a way for better detecting many-to-one and one-to-many correspondences. Our starting point is the set of alignments delivered by Moore's tool. As discussed above, these alignments have a very high precision, at the expense of an unsatisfactory recall. Our sentence alignment method considers these sentence pairs as being parallel and uses them to train a binary classifier for detecting parallel sentences. Using the predictions of this tool, it then attempts to align the remaining portions of the bitext (see Figure 2).

In Figure 2, Moore's links are displayed with solid lines; these lines delineate parallel pseudo-paragraphs in the bitexts (appearing in boxed areas), which we will try to further decompose. Note that two configurations need to be distinguished: (i) one side of a paragraph is empty: no further analysis is performed and a 0-to-many alignment is output; (ii) both sides of a paragraph are non-empty and define a i-to-j alignment that will be processed by the block alignment algorithm described below.

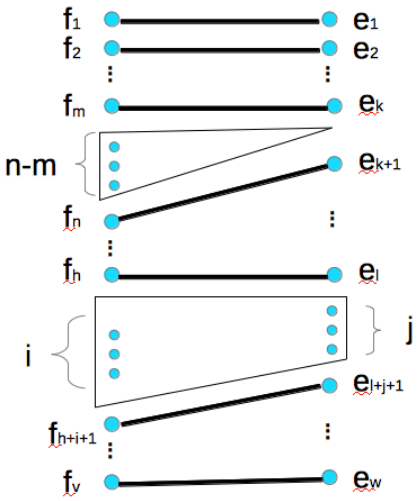

Figure 2: Filling alignment gaps

3.1 Detecting parallelism

Assuming the availability of a set of example parallel sentences, the first step of our approach consists in training a function for scoring candidate alignments. Following (Munteanu and Marcu, 2005), we train a Maximum Entropy classifier[9] (Rathnaparkhi, 1998); in principle, many other binary classifiers would be possible here. Our motivation for using a maxent approach was to obtain, for each possible pair of sentences ($\mathbf{f}$,$\mathbf{e}$), a link posterior probability $P(link|\mathbf{f},\mathbf{e})$.

We take the sentence alignments of the first step as positive examples. Negative examples are artificially generated as follows: for all pairs of positive instances $(\mathbf{e},\mathbf{f})$ and $(\mathbf{e}',\mathbf{f}')$ such that $\mathbf{e}'$ immediately follows $\mathbf{e}$, we select the pair $(\mathbf{e},\mathbf{f}')$ as a negative example. 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.

Formally, the 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:

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

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

- **lexical features**: for each pair of words[10] (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 number 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.

- **cognate features**: we loosely define cognates[11] as words sharing a common prefix of length at least 3. This gives rise to 4 features, which are respectively activated when the number of cognates in the parallel sentence is 0, 1, 2, or greater than 2.

- **copy features**: an extreme case of similarity is when a word is copied verbatim from the source to the target. This happens with proper nouns, dates, etc. We again derive 4 features, depending on whether the number of identical words in $\mathbf{f}$ and $\mathbf{e}$ is 0, 1, 2 or greater than 2.

[9] Using the implementation available from http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html.

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

[11] 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*.

3.2 Filling alignment gaps

The third step uses the posterior alignment probabilities computed in the second step to fill the gaps in the first pass alignment. 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[12].

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:

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

where $\mathbf{e}_{i':j'}$ is obtained by concatenation of all the sentences in the range $[i':j']$, and $S(i, j, k, l) = (j - i + 1)(l - k + 1) - 1$ is proportional to the block size. The factor $\alpha S(i', j', k', l')$ aims at penalizing large blocks, which, for the sentence-based metrics, yield much more errors than the small ones. This strategy implies to compute $O(|j - i + 1|^2 \times |k - l + 1|^2)$ probabilities, which, given the typical size of these blocks (see above), can be performed very quickly.

These values are then iteratively visited by decreasing order in a greedy fashion. The top-scoring block $i' : j', k' : l'$ is retained in the final alignment; all overlapping blocks are subsequently 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 in Figure 3: assuming that the best matching link is $\mathbf{f}_2$-$\mathbf{e}_2$, we delete all the links that include $\mathbf{f}_2$ or $\mathbf{e}_2$, as well as links that would imply a reordering of sentences, meaning that we also delete links such as $\mathbf{f}_1$-$\mathbf{e}_3$.

3.3 Experiments

In this section, we report the results of experiments run using again Jules Verne's book from the BAF corpus. Figures are reported in Table 3 where we contrast our approach with two simple baselines: (i) keep only Moore's links; (ii) complete Moore's links with one single many-to-many alignment for each block. For the maxent-based approach, we also report the precision on just those links that are not predicted by Moore. A more complete set of experiments conducted with other portions of the BAF are reported elsewhere (Yu et al., 2012) and have shown to deliver state-of-the-art results.

As expected, complementing the very accurate prediction of Moore's systems with our links significantly boosts the sentence-based alignment performance: recall rises from 0.62 to 0.80 for $\alpha = 0$, which has a clear effect on the corresponding F-measure (from 0.76 to 0.86). The performance differences with the default strategy of keeping those blocks unsegmented are also very clear. Sentence-wise, maxent-based alignments are also quite precise, especially when the value of α is chosen with care (P=0.91 for α=0.06); however, this optimization has a very small overall effect, given that only a limited number of alignment links are actually computed by the maxent classifier.

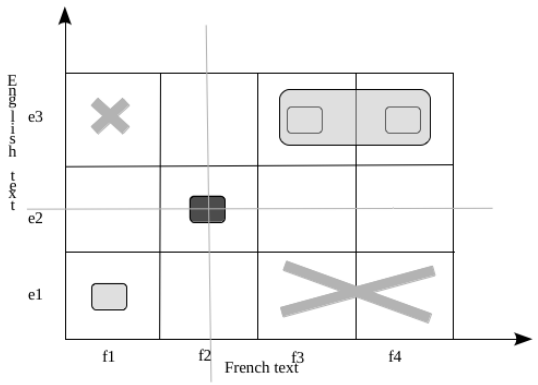

Figure 3: Greedy alignment search

	P		R	F
	(maxent)	(all)	(all)	(all)
link based				
Moore only	-	0.85	0.65	0.74
Moore+all links	-	0.78	0.75	0.76
Maxent, $\alpha = 0$	0.44	0.74	0.81	0.77
Maxent, $\alpha = 0.06$	0.42	0.72	0.82	0.77
sentence based				
Moore only	-	0.98	0.62	0.76
Moore+all links	-	0.61	0.88	0.72
Maxent, $\alpha = 0$	0.80	0.93	0.80	0.86
Maxent, $\alpha = 0.06$	0.91	0.97	0.79	0.87

Table 3: Performance of maxent-based alignments

[12] We enclose the source and target texts between begin and end markers to enforce alignment of the first and last sentences.

4 Sentence alignment in the real world

In this section, we analyze the performance obtained with our combined system, using excerpts of our small corpus as test set. For this experiment, the first two to three hundreds sentences in each book, corresponding to approximately two chapters, were manually aligned (by one annotator), using the same guidelines that were used for annotating the BAF corpus. Except for two books (EM and VF), producing these manual alignments was found to be quite straightforward. Results are in Table 4.

A first comment is that both baselines are significantly outperformed by our algorithm for almost all conditions and books. For several books (LM, AM, SW), the obtained sentence alignments are almost as precise as those predicted by Moore and have a much higher recall, resulting in very good overall alignments. The situation is, of course, much less satisfactory for other books (EM, VF, 5S). All in all, our method salvages many useful sentence pairs that would otherwise be left unaligned.

Moore's method remains remarkably accurate throughout the whole collection, even for the most difficult books. It also outputs a significant proportion of wrong links, which, for lack of reliable confidence estimators, are difficult to spot and contribute to introduce noise into the maxent training set.

The variation of performance can mostly be attributed to idiosyncrasies in the translation. For instance, *Emma* (EM) seems very difficult to align, which can be attributed to the use of an old translation dating back to 1910 (by P. de Puliga), and which often looks more like an adaptation than a translation. Some passages even question the possibility of producing any sensible (human) alignment between source and target[13]:

(en) *Her sister, though comparatively but little removed by matrimony, being settled in London, only sixteen miles off, was much beyond her daily reach; and many a long October and November evening must be struggled through at Hartfield, before Christmas brought the next visit from Isabella and her husband, and their little children, to fill the house, and give her pleasant society again.*

(fr) *La sœur d'Emma habitait Londres depuis son mariage, c'est-à-dire, en réalité, à peu de distance; elle se trouvait néanmoins hors de sa portée journalière, et bien des longues soirées d'automne devraient être passées solitairement à Hartfield avant que Noël n'amenât la visite d'Isabelle et de son mari.*

Les confessions (CO) is much most faithful to the content, yet, the translator has significantly departed from Rousseau's style[14], mostly made up of short sentences, and it is often the case that several French sentences align with one single English sentence, which is detrimental to Moore, and by ricochet, to the quality of maxent predictions. A typical excerpt:

(fr) *Pendant deux ans entiers je ne fus ni témoin ni victime d'un sentiment violent. Tout nourrissait dans mon coeur les dispositions qu'il reçut de la nature.*

(en) *Everything contributed to strengthen those propensities which nature had implanted in my breast, and during the two years I was neither the victim nor witness of any violent emotions.*

The same goes for Thackeray (VF), with a lot of restructurations of the sentences as demonstrated by the uneven number of sentences on both sides of the bitext. *Lord Jim* (LJ) poses another type of difficulty: approximately 100 sentences are missing on the French side, the rest of the text being fairly parallel (more than 82% of the reference links are actually 1-to-1). *Du côté de chez Swann* (SW) represents the other extreme of the spectrum, where the translation sticks as much as possible to the very peculiar style of Proust: nearly 90% of the reference alignments are 1-to-1, which explains the very good F-measure for this book.

It is difficult to analyze more precisely our errors; however, a fairly typical pattern is the inference of a 1-to-1 link rather than a 2-to-1 link made up of a short and a long sentence. An example from Hugo (TM), where our approach prefers to leave the second English sentence unaligned, even though the corresponding segment (*un enfant...*) is the in French sentence:

(fr) *Dans tout le tronçon de route qui sépare la première tour de la seconde tour, il n'y avait que trois passants, un enfant, un homme et une femme.*

(en) *Throughout that portion of the highway which separates the first from the second tower, only three foot-passengers could be seen. These were a child, a man, and a woman.*

A possible walk around for this problem would be to also add a penalty for null alignments.

[13] In this excerpt, in addition to several approximations, the end of the last sentence (*and their children...*) is not translated in French.

[14] Compare the number of sentences in Table 1.

				Moore				*Hunalign*		*Moore+maxent*				
				links	P	R	F	links	F	$S \neq 0$	$S = 0$	P	R	F
	fr	en	links	*link based*										
EM	160	217	164	84	0.76	0.39	0.52	173	0.43	72	10	0.52	0.53	0.52
JE	229	205	174	104	0.86	0.51	0.64	198	0.40	95	5	0.64	0.75	0.69
LM	232	205	197	153	0.97	0.76	0.85	203	0.63	64	2	0.79	0.87	0.83
LJ	580	682	515	403	0.94	0.73	0.82	616	0.60	155	15	0.82	0.81	0.76
VF	321	248	219	129	0.92	0.54	0.68	251	0.39	133	3	0.58	0.70	0.63
CO	326	236	213	104	0.86	0.42	0.56	256	0.28	135	3	0.62	0.70	0.66
5S	182	201	153	107	0.76	0.53	0.62	165	0.52	72	10	0.60	0.74	0.66
AM	258	226	222	179	1.00	0.81	0.90	222	0.71	55	0	0.88	0.93	0.90
TM	404	388	358	284	0.89	0.71	0.79	374	0.69	86	16	0.79	0.85	0.82
SW	492	495	463	431	0.94	0.87	0.90	474	0.80	59	9	0.85	0.92	0.88
	fr	en	links	*sentence based*										
EM	160	217	206	84	0.85	0.34	0.49	199	0.60	124	0	0.62	0.63	0.62
JE	229	205	270	104	0.92	0.36	0.52	235	0.60	125	0	0.90	0.76	0.82
LM	232	205	238	153	0.99	0.64	0.78	234	0.79	62	0	0.97	0.88	0.92
LJ	580	682	645	403	0.96	0.60	0.74	625	0.78	212	0	0.85	0.81	0.83
VF	321	248	363	129	0.98	0.35	0.52	318	0.62	163	0	0.88	0.71	0.79
CO	326	236	380	104	0.94	0.26	0.41	306	0.48	226	0	0.88	0.76	0.82
5S	182	201	260	107	0.98	0.40	0.57	224	0.70	81	0	0.93	0.67	0.78
AM	258	226	264	179	1.00	0.68	0.81	262	0.84	72	0	0.98	0.94	0.96
TM	404	388	445	284	0.96	0.61	0.75	418	0.82	134	0	0.93	0.87	0.90
SW	492	495	532	431	0.99	0.80	0.88	512	0.88	55	0	0.99	0.90	0.94

Table 4: Evaluating alignment systems on a sample of "real-world" books
For each book, we report the number of French and English test sentences, the number of reference links and standard performance measures. For the maxent approach, we also report separately the number of empty ($S = 0$) and non-empty ($S \neq 0$) paragraphs.

5 Conclusions and future work

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 able to significantly improve the bare results obtained with the sole Moore alignment system. Our study shows that the problem of sentence alignment for literary texts is far from being solved and additional work is needed to obtain alignments that could be used in real applications, such as bilingual reading aids.

The maxent-based approach proposed here is thus only a first step, and we intend to explore various extensions: an obvious way to go is to use more resources (larger training corpora, bilingual dictionaries, etc.) and add more features, such as part-of-speech, lemmas, or alignment features as was done in (Munteanu and Marcu, 2005). We also 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.

Regarding applications, a next step will be to produce and evaluate sentence alignments for a much larger and more diverse set of books, comprising more than 100 novels, containing books in 7 languages (French, English, Spanish, Italian, German, Russian, Portuguese) from various origins. Most were collected on the Internet from Gutenberg, wikisource and GoogleBooks[15], and some were collected in the course of the Carmel project (Kraif et al., 2007). A number of these books are translated in more than one language, and some are raw OCR outputs and have not been cleaned from errors.

Acknowledgments

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

[15] http://books.google.com

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.

William A. Gale 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.

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

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

Olivier Kraif, Marc El-Bèze, Régis Meyer, and Claude Richard. 2007. Le corpus Carmel: un corpus multilingue de récits de voyages. In *Proceedings of Teaching and Language Corpora : TaLC'200*, Paris.

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.

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.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In Stephen D. Richardson, editor, *Proceedings of the annual meeting of tha Association for Machine Translation in the Americas (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.

John Nerbonne, 2000. *Parallel Texts in Computer-Assisted Language Learning*, chapter 15, pages 354–369. Text Speech and Language Technology Series. Kluwer Academic Publishers.

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

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, Granada, Spain.

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

Jakob Uszkoreit, 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.

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

Qian Yu, Aurélien Max, and François Yvon. 2012. Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *Proceedings of the Language Resource and Evaluation Conference (LREC)*, Istambul, Turkey.