An Iterative Approach for Mining Parallel Sentences in a Comparable Corpus

Lise Rebout, Philippe Langlais
DIRO
Université de Montréal
CP 6128 Succursale Centre-Ville
H3C3J7 Montreal, Québec, Canada
lise.rebout@gmail.com, felipe@iro.umontreal.ca

Abstract

We describe an approach for mining parallel sentences in a collection of documents in two languages. While several approaches have been proposed for doing so, our proposal differs in several respects. First, we use a document level classifier in order to focus on potentially fruitful document pairs, an understudied approach. We show that mining less, but more parallel documents can lead to better gains in machine translation. Second, we compare different strategies for post-processing the output of a classifier trained to recognize parallel sentences. Last, we report a simple bootstrapping experiment which shows that promising sentence pairs extracted in a first stage can help to mine new sentence pairs in a second stage. We applied our approach on the English-French *Wikipedia*. Gains of a statistical machine translation (SMT) engine are analyzed along different test sets.

Keywords: comparable corpora, nearly-parallel document mining, machine translation

1. Introduction

This work is concerned with mining parallel sentences from a comparable corpus. A typical two-stage approach for tackling the problem consists in identifying a set of potentially comparable documents, then to mine parallel sentences among those document pairs. The first stage is often done heuristically. In the news domain, many authors have for instance proposed to rule out news written at too different times; *e.g.* (Munteanu and Marcu, 2005). Other constraints, such as length ratio can further restrain the document pairs considered. Others, working with *Wikipedia* avoid document pairing and rely instead on the inter-lingual links present in this resource (Adafre and de Rijke, 2006; Smith et al., 2010).

There are notable exceptions to this trend of work. In (Tillmann, 2009) the author avoids document pairing by directly harvesting the cartesian product of sentences in the source and target collections. This represents an enormous space trough which he manages to search thanks to a smart organization of the operations involved in computing the probability given by an IBM Model 1 to any sentence pair. Thus, the approach is specific to this feature only. On the contrary, the work of (Ture et al., 2011) tackles the computationally challenging task of pairwise document comparison thanks to a map-reduce approach which — provided enough computers (they used 96 cores) — was shown to be efficient for pairing all the German-English articles in *Wikipedia*. In (Ture and Lin, 2012) the authors describe the pending experiment, where pairwise sentence comparison is conducted over the huge set of comparable documents they identified (over 64 million document pairs), again making efficient use of map-reduce and a reasonable sized cluster.

Although the last three studies we discussed are impressive engineering success stories, we take a rather opposite direction, guided by the intuition that mining parallel sentences from a set of *nearly-parallel* document pairs (a notion we define later on) will likely yield to better parallel sentence pairs than mining all (or many) document pairs, as done in (Ture and Lin, 2012). We see several reasons why it is sensible to focus of nearly-parallel documents in the first place. From a computational point of view, it is obviously interesting to limit the number of documents among which pairs of sentences will be searched for; regardless of the fact that the heavy computations can be smartly parallelized. Another argument in favour of measuring the parallelness of a document pair is that it might help in adapting the technology with which to extract parallel sentences. For instance, we might prefer to extract parallel sentences of very parallel documents thanks to a standard sentence alignment technique (Gale and Church, 1993) instead of using a classifier, as typically done.

We present our approach to mine parallel sentences in a comparable corpus in section 2. We describe the resources we used for training our classifiers, and those we used for conducting our SMT experiments in section 3. We report a number of experiments we conducted in section 4. We discuss our work and present future avenues in section 5.

2. Approach

Our approach has two main components. The first one selects fruitful pairs of documents in a large set of document pairs, that is, pairs of documents that are likely parallel. The second component is taking care of the pairwise sentence comparison conducted for each identified pair of documents. A block diagram of the system with references to the sections that discuss each component is provided in Figure 1.

2.1. Parallel Document Mining

We compare two approaches for identifying useful pairs of documents. The first one relies on a classifier trained in a supervised way. The second one is a cross-lingual information retrieval (IR) system similar to (Utiyama and Isahara, 2003). Both methods are detailed in the following sections.

2.1.1. Classifier

We want to estimate the parallelness of a pair of documents. Instead of addressing this regression problem head on, we

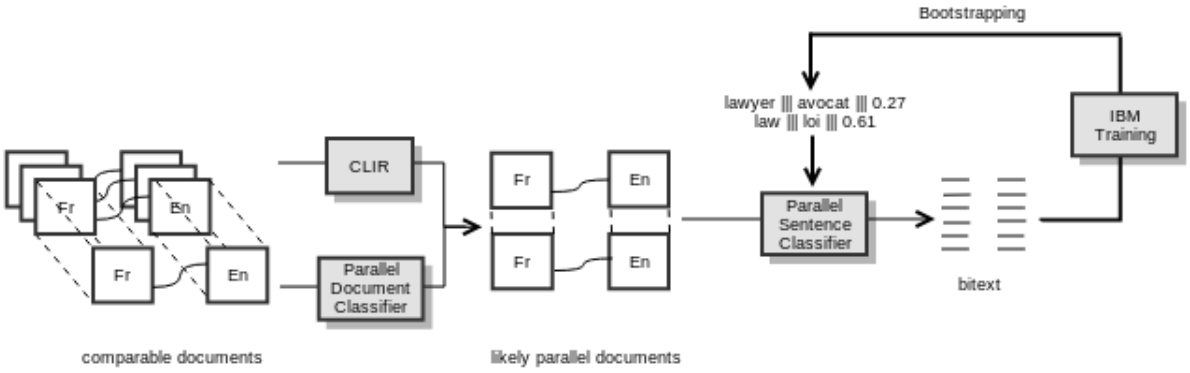

Figure 1: Block diagram of the approach

recast it into a binary classification problem, where the task is to define nearly-parallel document pairs from others. We measure the degree of parallelness of a sentence-aligned document pair of n_{src} source and n_{trg} target sentences with the ratio:

$$r_{para} = \frac{2 \times n_{para}}{n_{src} + n_{trg}}$$

where n_{para} designates the number of parallel sentences. Based on it, we arbitrarily define the class of *nearly-parallel* document pairs as those which ratio r_{para} is greater or equal to 2/3.

We trained in a supervised way a classifier to recognize this class, thanks to an annotated corpus described in section 3.2. In order to do so, we compute a handful of lightweight features, making use of a dictionary extracted from the titles of the article pairs linked in *Wikipedia*, and mainly following the work of (Adafre and de Rijke, 2006). In particular, the following features are computed for each pair of documents, and are illustrated in Figure 2:

- the number of sentences in the French and the English articles,
- the Levenshtein distance between the phrases collected from the wiki links present in the English articles (wiki anchors), and those collected in the French articles, then translated (whenever possible) thanks to the WIKITITLE corpus. For instance, in Figure 2 the French phrase `loi fondamentale du royaume des pays bas` is translated into the English one `constitution of the netherlands`,
- the number of phrases in each article that match an entry in WIKITITLE, but that are not anchored in wiki links (extended links); the size of the intersection of those two sets of phrases, after the French phrases got translated thanks to WIKITITLE.

Those features are definitely *Wikipedia*-centric, therefore our parallel document classifier is currently specific to this collection of documents. To us, WIKITITLE is playing the role of a dedicated bilingual lexicon. It remains to investigate whether a general bilingual lexicon would be as adequate or even better.

2.1.2. Cross-lingual IR System

We implemented a cross-lingual information retrieval approach very similar to the one described in (Utiyama and Isahara, 2003). The idea is to index each English document thanks to a vector space model built on the English words present in the collection. At retrieval time, we "translate" a French document, so that it can be represented in the same vector space. We used an in-house bilingual dictionary for performing the translation which contains a total of 29 861 different forms[1] and used the LUCENE API[2] in its default setting for the indexing and retrieval steps.

In our implementation, English documents with a similarity score over 0.5 were considered nearly-parallel to the queried French document, provided they defined a pair belonging to the set of interlingually linked pairs of articles. The value of 0.5 was chosen empirically so that the quantity of mined document pairs with LUCENE was directly comparable to the number of pairs mined by the classifier.

2.2. Parallel Sentence Mining

Once the nearly-parallel documents are identified, we harvest all possible sentence pairs and select those that are parallel according to a classifier trained in a supervised way (see section 4.2.1.). This is a very imbalanced classification task where the number of negative examples is quadratic in the number of sentences, while the number of parallel sentences is at best linear. Lexical overlap filtering strategies, as in (Munteanu and Marcu, 2005) could be used for reducing computation time.

We computed over 20 features of various nature including most of those described in (Smith et al., 2010), some we adapted from (Adafre and de Rijke, 2006), and others we specifically engineered. Some features are characterizing a sentence by itself (e.g. number of words or characters), a sentence within a document (e.g. the presence of hapax word, which has been shown to be useful for aligning documents (Enright and Kondrak, 2007)), a pair of sentences (such as the number of links or numbers they share, Jaccard coefficients comparing different representations of those sentences, etc.).

[1] Multiple translations of a given word were all considered.
[2] `http://lucene.apache.org/core/`

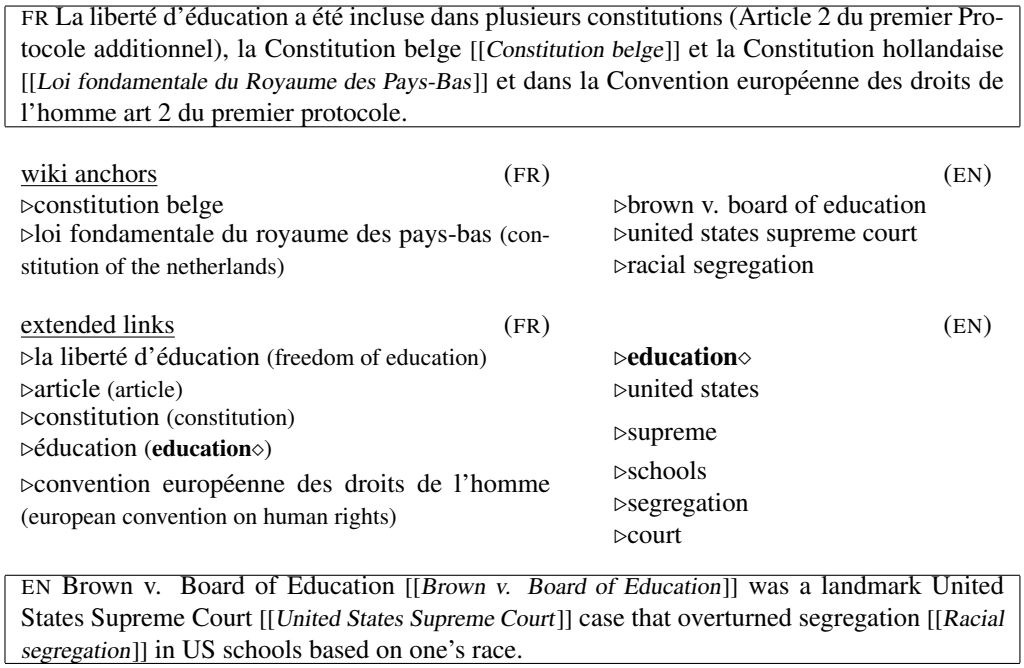

Figure 2: Illustration of the representation of two *Wikipedia* documents, according to the text anchored in wiki links, or the phrases that correspond to article titles in *Wikipedia*. When present, the translation of each French phrase in WIKITITLE is reported in parenthesis. The only match in both representations is marked by a diamond symbol.

2.2.1. Post-treatment

t As we will show, the identification of parallel sentence pairs is error prone. This is why we investigated a number of post-treatment strategies. The simplest one imposes that a pair of sentences receives a classification score higher than a given threshold ρ in order to get elected parallel. Still, the main reason for the noise we observe is that the decisions made by the classifier are done independently for each sentence pair. This may lead to situations where, for instance, a given French sentence is paired to several target sentences.[3]

The set of pairings made at given threshold ρ can be organized into a bipartite graph where nodes represent French and English sentences, and edges represent French to English sentence pairings and are labeled with the cost of the pairing as returned by the classifier. Refining the decisions made can then be casted into finding the matching (that is, a set of edges that do not share vertices) with the maximum cost, where the cost of a match is the sum of the cost of all the pairs (edges) involved. We implemented two well known solutions to this problem, a greedy maximum first approach (hereafter named *greedy*) and the so-called Hungarian algorithm (hereafter named *hung*). The former picks the candidate edge e with the maximum score, removes candidate edges that share a vertice with e, and iterates until no edge remains.

Last, none of the post-processing methods aforementioned guaranty the sequentiality of the pairs we typically observe in a document. Therefore, we implemented a last heuristic, hereafter named *extend*, which adds the edge $(s+1, t+1)$ whenever it does not share a vertices with existing edges, its score by the classifier is at least positive (but possibly lower than ρ) and edges (s, t) and $(s+2, t+2)$ are already selected. This heuristic which admittedly is rather specific proved to be useful, as shown in section 4.

2.3. Bootstrapping

None of the features we described so far are exploiting a general bilingual dictionary. The main motivation for this was that generic dictionaries might be of little use in mining domain specific comparable corpora, an assumption we still need to assess.[4] It also makes our approach more applicable to language pairs for which covering dictionaries are not available. Still, using a dictionary has been reported to be useful (Munteanu and Marcu, 2005; Smith et al., 2010; Ture and Lin, 2012). This motivates an iterative approach, where a first pass extracts good sentence pairs, from which a bilingual dictionary is inferred. This dictionary can then be used in successive passes of pairing. This is similar in spirit to the 2-stage approach proposed in (Ture and Lin, 2012) where a fast classifier is first applied, producing a large set of sentence pairs which is refined by a more accurate but slower classifier. One difference is that in our case, the second classifier is bootstrapped by the knowledge acquired from the first stage. We investigate one such iteration in section 4.

3. Resources

3.1. Comparable Corpus

We downloaded the May 2011 dumps of the English and the French parts of *Wikipedia*. The English part contains 3.7

[3]Note that this might indicate a legitimate 1-to-many alignment, a situation we do not consider in this work.

[4]See the ACCURAT project on this matter `http://www.accurat-project.eu`.

millions of articles, the French part contains 1.1 millions of articles. A total of 551 388 pairs of articles are marked by an interlingual (English-French) link. This quantity of data is a mid-ground between very large data collections such as the *GigaWord* corpus (which English version contains roughly 26 gigabytes of texts) and more modest corpora such as the TDT3 corpus used in (Fung and Cheung, 2004), and which English part contains 290 000 sentences. We parsed the XML format of the dumps thanks to the SAX API for Java. We removed the specific wiki markup, keeping the record of useful information such as wiki hyperlinks. The resulting text was then segmented and tokenized. We removed all articles containing less than 10 sentences, which is motivated by computation time considerations, as well as by the fact that short documents typically contain sentences specific to *Wikipedia* (such as `References` or `External links`) as well as very repetitive constructions. In the end, we collected 367 797 pairs of articles linked by an interlanguage link. See Table 1 for details.

	EN	FR
sentences	33 028k	24 331k
words	70 393k	304 445k
avg. #sent. per article	59.9	43.1
avg. #words per sentence	14.3	12.5
#pairs of linked articles	551k	
#pairs of articles considered	367 797	

Table 1: Main characteristics of the inter-lingually linked articles in the English-French *Wikipedia* downloaded in May 2011.

3.2. Manually Annotated Corpus

For the sake of training our classifiers, we manually annotated a (random) subset of 80 article pairs. Naturally, this set is disjoint from the *Wikipedia* material described in the previous section. For each French sentence, we marked its English translation in the corresponding English document, if any. Following (Smith et al., 2010), we consider as translation, literally translated sentences. A total of 2 057 parallel sentence pairs where manually identified. Out of the 80 pairs of articles we annotated, we found that 17 (21.3%) were nearly-parallel, as defined in section 2.1.1.

As far as the of May 2011 *Wikipedia* dump is concerned, the English article `Nangchen horse` is for instance considered to be in translation relation with its French counterpart, while the biography of `Albin Egger-Lienz` is not. The annotation, although simple in nature, turned out to be particularly time consuming and is available for download.

3.3. SMT data sets

The in-domain training material we used in this study is made of the Europarl and the news commentary training sets available from the WMT 2011 evaluation campaign. It gathers a total of 1 940 639 sentence pairs and is named EUROPARL hereafter.

We built a development corpus of 2 489 sentence pairs from the news dev set available from WMT 2011, plus 813 sentence pairs we extracted from *Wikipedia* and which parallelness was verified manually.

We gathered three different test sets. One, named NEWSTEST, is the news test set of WMT 2011 and contains 3003 test sentences. The EUROTEST corpus is composed of 2 000 sentence pairs from the European parliament corpus that we took from the WMT 2008 test set. Last, we also gathered 800 parallel sentence pairs we collected from *Wikipedia*. As for the development set, we paid attention to verify that those sentences were absent from the comparable corpus we mined sentences from.

3.4. Bitexts Extracted

Out of the *Wikipedia* material we describe in Section 3.1., we automatically extracted three bitexts:

WIKITITLE we gathered the titles of *Wikipedia* articles that are interlingually linked

WIKICLASS the bitext collected by first running our (best) document-level classifier (see section 4.1.1.), then by applying the same (best) parallel-sentence classifier we devised (see section 4.2.1.)

WIKILUCENE was obtained by first running our IR system (see section 4.1.2.), then applying the same (best) sentence-pair classifier as previously (see section 4.2.1.)

The main characteristics of those bitexts are reported in Table 2. It is interesting to note that although WIKICLASS and WIKILUCENE have been extracted roughly from the same number of document pairs (see section 4.1.), the number of pairs of sentences extracted varies greatly among the two corpora: the documents retrieved by LUCENE are typically longer than the ones returned by our classifier.

	WIKITITLE		WIKICLASS		WIKILUCENE	
	FR	EN	FR	EN	FR	EN
sent.	580k		561k		1 454k	
words	1.5m	1.5m	11.1m	10.7m	29.4m	27.3m
types	346k	337k	399k	382k	665k	612k

Table 2: Main characteristics of the bitexts extracted from *Wikipedia*.

4. Experiments

In the two classification tasks we framed, we compare two families of classifiers, namely single-layer perceptrons and decision trees. The former has been chosen because it delivers state-of-the-art performance in numerous learning tasks, plus it includes as a special case the so-called maxent approach popular in several works on sentence pair extraction (Munteanu and Marcu, 2005). The latter has been used because of its simplicity as well as the interpretation power it offers. It is also known to be a good baseline. We investigated a number of meta-parameters that can influence the performance of those classifiers. For the perceptrons,

we varied the number of hidden units (from 10 to 1000), the number of hidden layers (0 or 1), the number of iterations (up to 10 000), as well as the learning rate and the momentum. For the decision trees, we varied the learning algorithms (RandomTree or J48). This exploration was facilitated by the WEKA toolkit (Hall et al., 2009).

We evaluate the classifiers by their ability to correctly detect the positive instances, in our case, the parallel (document or sentence) pairs. Therefore precision is computed as the percentage of identified parallel pairs that are truly parallel, and recall measures the percentage of truly parallel pairs that are correctly identified. We also report the F_1 combination of these two scores. The evaluation is conducted on the manually annotated ressource described in section 3.2., using a cross-validation procedure.

4.1. Parallel Document Mining

4.1.1. Classification

Table 3 reports the best results we obtained for each classifier family, as measured by F_1 after a 3-fold cross-validation procedure. Neural networks achieved the best performance and the best performing one was trained over 500 epochs with a learning rate of 0.9 and a momentum of 0.3. Somehow at a surprise, we found that using no hidden layer leads to better results. For the decision tree, the best variant was trained using the J48 algorithm.

	Recall	Precision	F_1-score
neural network	82.4%	87.5%	84.8%
decision tree	70.6%	75.0%	72.7%

Table 3: Performance of the two best configurations of document-based classifiers.

The best neural network configuration we observed was retrained on the full annotated material, and the resulting classifier was applied to the 367 797 pairs of *Wikipedia* articles we considered in this study. This resulted into 38 829 pairs of documents identified as nearly-parallel, that is, 10.5% of the pairs of documents. This is consistent with the observation made by (Patry and Langlais, 2011) who estimate that over 44 000 pairs of *Wikipedia* English-French pages of a 2009 *Wikipedia* dump are indeed parallel; a non negligible quantity.

4.1.2. Cross-lingual IR

We applied our cross-lingual IR system on the same *Wikipedia* corpus and retrieved 43 564 article pairs, a larger set than the one retrieved by our best classifier. The documents returned by LUCENE in each language contain roughly twice the number of sentences of the documents retrieved by the best classifier. This shows a bias of the IR approach toward larger documents, which might be due to a lack of coverage of our bilingual dictionary. It is also worth mentioning that only 12 490 article pairs have been returned by both approaches, which indicates that either the approaches are complementary, or one is suspiciously noisy. This will be analyzed in the section 4.2.2.

4.2. Parallel Sentence Mining

4.2.1. Classification

We used the 17 document pairs that were manually found nearly-parallel in the manually annotated corpus described in section 3.2. The results reported are averaged over a 17-fold cross-validation procedure (the test set being formed by one document at a time). The winning configuration is a neural network trained over 1 000 epochs with a learning rate of 0.2, a momentum set to 0.3 and a hidden layer of 16 neurones. The best decision tree configuration was obtained with the J48 training algorithm. The performance of those two configurations are reported in the two first lines of Table 4. Note that identifying parallel sentences is a more difficult task than identifying parallel documents and that the performance of the two winning configurations only differ by one absolute F_1 point.

We tested a number of refinement techniques presented in section 2.2.1. For both the decision tree and the multi-layer perceptron output, applying refinements helps increasing the performance significantly. This shows that taking into account previous decisions, as done in (Smith et al., 2010) is a good idea, as far as performance is concerned (on the bad stand, it incurs a computational cost). The winner configuration is marked by a $\diamond$ sign in Table 4 and corresponds to a neural network which decisions are post-processed by the hungarian algorithm, and the extension heuristic described in section 2.2.1. This configuration improves the neural network alone by more than 11 absolute F_1 points (from 56.3 to 67.6). The last line of the table is discussed in section 4.4.

		Rec.	Prec.	F_1
(nn)	neural network	57.5%	55.2%	56.3%
(dt)	decision tree	53.9%	56.9%	55.4%
Document-level post-processing ($\rho = 0.1$):				
	dt + *hung*+ *extend*	53.9%	69.6%	60.8%
	nn + *greedy*	60.5%	72.5%	66.0%
	nn + *hung*	58.3%	73.1%	64.9%
($\diamond$)	nn + *hung*+ *extend*	62.7%	74.4%	67.6%
Iteration:				
	$\diamond$ + bootstrapping	72.0%	80.2%	75.9%

Table 4: Performance of different configurations of sentence-based classifiers.

The best configuration ($\diamond$) was retrained on the full set of manually annotated sentence pairs (the 17 documents) before being applied to the documents pairs collected by the classifier (section 4.1.1.) or by LUCENE (section 4.1.2.). The main characteristics of the resulting bitexts (WIKICLASS and WIKILUCENE respectively) have been discussed in section 3.4. and are summarized in Table 2.

4.2.2. Manual Evaluation

A blind evaluation of a random excerpt of 200 pairs of sentences from WIKICLASS and from WIKILUCENE reveals that 70-76% of the sentence pairs sampled from WIKICLASS are indeed parallel or semi-parallel, while this is

true for only 18-22% of the sentence pairs we sampled in WIKILUCENE. See Table 5 for more details.
Clearly, filtering adequately the document pairs helps the sentence classifier to recognize parallel pairs. Indeed, we conducted an experiment on training a classifier with randomly selected pairs of interlingually linked articles in *Wikipedia* (regardless of their parallelness ratio r_{para}) and found that none of the sentence classifiers we trained could learn to recognize parallel sentences. Those figures, although measured on a small excerpt of sentence pairs, suggest that the quantity of sentences mined does not necessarily warrant quality, an observation which is in line with the work of (Morin et al., 2007) which questions, in the context of term-translation mining, the popular motto that more data is better data.

	WIKICLASS		WIKILUCENE	
parallel	70%	140	18.5%	37
semi-parallel	6%	12	4.0%	8
non-parallel	24%	48	77.5%	155

Table 5: Manual evaluation of two random excerpts of 200 sentence pairs randomly extracted from WIKICLASS and WIKILUCENE.

4.3. Machine Translation Evaluation

We trained a number of French-English SMT systems using the Moses toolkit (Koehn et al., 2007) in its default setting. Since our interest is to measure the quality of the parallel material acquired, we used the same language model for all the system configurations. This model, a Kneser-Ney 5-gram model was trained with the SRILM package (Stolcke, 2002) on the union of all the target material we collected (EUROPARL+WIKICLASS+WIKILUCENE). The BLEU metric as well as the sentence (SER) and the word (WER) error rates are reported in Table 6 for the three test sets described in section 3.3.
This table deserves some comments. First, for the three test sets, training on WIKICLASS plus EUROPARL leads to improvements in WER and BLEU over a system trained on one of these corpora only. According to BLEU, the best configuration on WIKITEST and NEWSTEST corresponds to the system trained on the concatenation of EUROPARL and WIKICLASS, which confirms the overall quality of the material we acquired. We note however that on EUROTEST, the bitext extracted by LUCENE leads to the overall best BLEU score. We do not have a clear explanation yet for why this is so. It is also interesting to note that on the WIKITEST task, the WIKICLASS training set alone yields to better performance than the out-domain EUROPARL training set.
Among the 3 003 sentences of the NEWSTEST set, the translations produced by the systems trained on EUROPARL and EUROPARL+WIKICLASS differ in as much as 2 517 cases. Figure 3 reports a few selected examples of diverging translations. Example (a) illustrates an example where the EUROPARL system translates the past tense in the present tense because this is the preferred tense in EUROPARL. Example (b) shows a case of a deceptive translation produced by the EUROPARL system, but corrected by the other. Most examples we analyzed reveal the better lexical choices made by the adapted engine.
Table 7 provides the percentage of unknown unigram and bigrams in the test sets. On the two out-domain test sets the percentage of unknown units remains high (around 20% for bigrams). On the EUROTEST translation task, those rates are much lower, thanks in great part to the large portion of in-domain training material.

	WIKITEST		NEWSTEST		EUROTEST	
$\neq$ words	12 172		4 309		7 561	
$\neq$ bigrams	49 031		12 271		32 988	
	1-g	2-g	1-g	2-g	1-g	2-g
EU	17.3	37.6	11.0	24.3	0.8	5.9
WC	8.9	31.2	11.0	33.9	6.5	31.3
EU+WT	10.0	35.1	6.5	23.4	0.6	5.9
EU+WC	7.4	25.5	5.8	19.9	0.5	5.5
EU+WC+WT	6.6	25.0	4.9	19.6	0.5	5.5
EU+WL+WT	5.3	21.1	4.3	17.1	0.4	5.1

Table 7: Percentages of out of vocabulary words and bigrams for different training sets. EU, WC, WL and WT stand respectively for EUROPARL, WIKICLASS, WIKILUCENE and WIKITITLE.

4.4. Bootstrapping

Since our document classifier is expected to spot documents with at least two thirds of the sentences being aligned, we kept the documents for which this ratio was actually met according to the sentence classifier. This represents 7 516 document pairs, that is, 19.7% of the documents our classifier identified as nearly-parallel in the first place. This material was used for training an IBM model 1 using the MGIZA++ toolkit (Gao and Vogel, 2008). We used this model for computing a number of lexical features that we added to the set of features already used by the sentence-level classifier. More precisely, and very similarly to (Munteanu and Marcu, 2005), we computed for a pair of English and French sentences (e, f):

- IBM model 1 estimate of $p(e|f) = \prod_{t \in e} \sum_{s \in f} p(t|s)$,
- min and max values of $\sum_{s \in f} p(t|s)$ over each t in e,
- number of French words associated to at least one word in the English sentence pair,
- number of French and English unknown words according to the translation model.

The last line of Table 4 reports the performance of the neural network classifier retrained with this extended feature set. An absolute 8% increase in F_1-score is observed upon the best classifier we trained during the first iteration. The last line of Table 6 reports the gains in translation obtained by this bootstrapping procedure. For all test sets, we observe an increase of BLEU and a decrease of the WER score, compared to the configuration

	WIKITEST			NEWSTEST			EUROTEST		
	WER	SER	BLEU	WER	SER	BLEU	WER	SER	BLEU
EUROPARL	75.67	95.56	10.44	58.32	99.17	21.54	53.93	97.60	27.51
WIKICLASS	73.54	94.68	12.04	60.19	99.50	20.10	61.34	99.10	20.43
EUROPARL+WIKITITLE	74.37	94.68	11.56	58.19	**99.13**	21.85	**53.74**	**97.50**	27.71
EUROPARL+WIKICLASS	73.34	94.55	**12.20**	**57.62**	99.20	**22.46**	53.83	97.70	27.71
EUROPARL+WIKICLASS+WIKITITLE	**73.17**	**94.17**	12.11	57.97	99.23	22.19	54.00	97.75	27.38
EUROPARL+WIKILUCENE+WIKITITLE	76.11	96.07	10.82	59.08	99.27	21.06	53.81	97.70	**28.12**
EUROPARL+BO+WIKITITLE	**72.98**	**94.42**	**12.38**	**57.58**	**99.13**	**22.56**	**53.71**	97.60	27.65

Table 6: Machine translation performance on two different test sets, as a function of the training set (first column).

EUROPARL+WIKICLASS+WIKITITLE. This indicates that bootstrapping is a fruitful strategy that deserves further investigations, which is left as future work.

5. Conclusion and Perspectives

We have described a simple implementation of an iterative two-stage approach for mining parallel sentences in a comparable corpus. This approach has the interesting property of being endogenous in the sense that it does not assume the existence of a large external bilingual dictionary,[5] as most existing approaches do. We provide some evidence that filtering the set of pairs of documents among which to extract parallel sentences has several potential advantages, among which better parallel material collected as measured manually and on two machine translation test sets. We developed a classifier for recognizing nearly-parallel document pairs, that is, documents in which at least two thirds of the sentence pairs are parallel. We show that it outperforms the standard cross-lingual information retrieval approach of (Utiyama and Isahara, 2003) which makes uses of a general bilingual dictionary. This in turn leads to better sentence pair extraction, as reflected by the SMT experiments we conducted. We also implemented a first iteration of bootstrapping which leads to systematic improvements in sentence pair classification, as well as in SMT.

This work opens up a number of avenues. First, the parallel document classifier we investigated here makes use of features tailored to *Wikipedia*. It remains to see whether this could be deployed on other collections of documents, possibly exploiting a general bilingual dictionary. Second, we must investigate the portability of our approach to other pairs of languages. One interesting point to look at is wether the document- and sentence-level classifiers we trained for the English-French language pair could be reused for other language pairs, as done in (Smith et al., 2010). Also, we accomplished a very simple bootstrapping iteration in this study. We need to investigate this further, notably by verifying whether the document-level classifier can benefit this iteration as well, especially since the number of training examples given to it in this study is very low. Measuring the number of iterations that can be accomplished successfully is another aspect we want to analyze.

In this study, we considered the classification of document pairs as a binary decision process (parallel or not). We plan to investigate if a finer level of granularity can be learned. Another trend of work we would like to pursue is to measure the effectiveness of our approach for tackling domain specific comparable corpora (possibly simulated by the *Wikipedia* categories). This is a useful setting for adapting an SMT engine to a new domain where general bilingual dictionaries might not be very useful.

The material we gathered in this study is available for download at `http://rali.iro.umontreal.ca/rali/?q=fr/node/1293`.

[5] It does rely on a dictionary built up from the titles of the interlingually linked article in *Wikipedia*.

Acknowledgments

This work has been partly founded by Natural Sciences and Engineering Research Council of Canada.

6. References

Adafre, S. and de Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, page 62–69.

Enright, J. and Kondrak, G. (2007). A Fast Method for Parallel Document Identification. In *NAACL HLT 2007, Companion Volume*, pages 29–32, Rochester, NY.

Fung, P. and Cheung, P. (2004). Multi-level bootstrapping for extracting parallel sentences from a quasi-comparable corpus. In *Proceedings of the 20th international conference on Computational Linguistics*.

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

Gao, Q. and Vogel, S. (2008). Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, Issue 1(10–18).

(a)	source	Il **était** frappant de constater **combien** la question de la **partie continentale** elle-même était absente.
	reference	It was striking how the issue of the mainland itself was absent.
	EUROPARL	It **is** striking **that** the issue of the **continent** itself was absent.
	EU+WC+WT	It **was** striking **how** the issue of the **mainland** itself was absent.
(b)	source	C'est le serveur de la BBC **qui a transmis** cette information.
	reference	The case was reported by BBC.
	EUROPARL	This is the BBC server **which has received** this information.
	EU+WC+WT	This is the BBC server **which reported** this information.
(c)	source	Zapatero et la **ligne rouge** allemande
	reference	Zapatero and the German red line
	EUROPARL	Zapatero and the German **Red**
	EU+WC+WT	Zapatero and the German **red line**
(d)	source	Mystérieux **travaux de terrassement**
	reference	Mysterious **earthworks**
	EUROPARL	Mysterious **workings of terrassement**
	EU+WC+WT	Mysterious **earthworks**
(e)	source	Elle a écrit le **scénario** elle-même: c'est une **histoire d'amour** entre une femme originaire de la Bosnie et un homme d'origine **serbe**.
	reference	She wrote the script herself - a love story between a woman from Bosnia and a Serbian man.
	EUROPARL	She wrote the **scenario** itself: It is a **history of love** between a woman from Bosnia and a man of **Serb** origin.
	EU+WC+WT	She wrote the **screenplay** itself: It is a **love story** between a woman from Bosnia and a man of **Serbian** origin.

Figure 3: Selected translation examples produced on the NEWSTEST data set. Sentences have been detokenized manually for the sake of readability. EU+WC+WT is a shortcut for the bitext obtained by concatenating EUROPARL, WIKICLASS and WIKITITLE.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL, demo session*, Prague, Czech Republic.

Morin, E., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *45th ACL*, pages 664–671, Prague, Czech Republic, June.

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

Patry, A. and Langlais, P. (2011). Identifying parallel documents from a large bilingual collection of texts: application to parallel article extraction in wikipedia. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 87–95.

Smith, J., Quirk, C., and Toutanova, K. (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*, page 403–411.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of ICSLP*, Denver, Colorado, Sept.

Tillmann, C. (2009). A Beam-Search extraction algorithm for comparable data. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, page 225–228.

Ture, F. and Lin, J. (2012). Why not grab a free lunch? mining large corpora for parallel sentences to improve translation modeling. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 626–630, Montréal, Canada, June.

Ture, F., Elsayed, T., and Lin, J. (2011). No free lunch: brute force vs. locality-sensitive hashing for cross-lingual pairwise similarity. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 943–952.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, page 72–79.