Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia.

Alexandre Patry
KeaText
845, Boulevard Dcarie, bureau 202
Saint-Laurent, Canada H4L 3L7
alexandre.patry@keatext.com

Philippe Langlais
DIRO/RALI
Université de Montréal
Montréal, Canada H3C3J7
felipe@iro.umontreal.ca

Abstract

While several recent works on dealing with large bilingual collections of texts, *e.g.* (Smith et al., 2010), seek for extracting **parallel sentences** from comparable corpora, we present PARADOCS, a system designed to recognize pairs of **parallel documents** in a (large) bilingual collection of texts. We show that this system outperforms a fair baseline (Enright and Kondrak, 2007) in a number of controlled tasks. We applied it on the French-English cross-language linked article pairs of Wikipedia in order see whether parallel articles in this resource are available, and if our system is able to locate them. According to some manual evaluation we conducted, a fourth of the article pairs in Wikipedia are indeed in translation relation, and PARADOCS identifies parallel or noisy parallel article pairs with a precision of 80%.

1 Introduction

There is a growing interest within the Machine Translation (MT) community to investigate *comparable corpora*. The idea that they are available in a much larger quantity certainly contributes to foster this interest. Still, *parallel corpora* are playing a crucial role in MT. This is therefore not surprising that the number of *bitexts* available to the community is increasing.

Callison-Burch et al. (2009) mined from institutional websites the 10^9 word parallel corpus[1] which gathers 22 million pairs of (likely parallel) French-English sentences. Tiedemann (2009) created the Opus corpus,[2] an open source parallel corpus gathering texts of various sources, in several languages pairs. This is an ongoing effort currently gathering more than 13 Gigabytes of compressed files. The Europarl corpus[3] (Koehn, 2005) gathers no less than 2 Gigabytes of compressed documents in 20 language pairs. Some other bitexts are more marginal in nature. For instance, the novel *1984* of George Orwel has been organized into an English-Norvegian bitext (Erjavec, 2004) and *Beyaz Kale* of Orhan Pamuk as well as *Sofies Verden* of Jostein Gaardner are available for the Swedish-Turk language pair (Megyesi et al., 2006).

A growing number of studies investigate the extraction of near parallel material (mostly sentences) from comparable data. Among them, Munteanu et al. (2004) demonstrate that a classifier can be trained to recognize parallel sentences in comparable corpora mined from news collections. A number of related studies (see section 5) have also been proposed; some of them seeking to extract parallel sentences from cross-language linked article pairs in Wikipedia[4] (Adafre and de Rijke, 2006; Smith et al., 2010). None of these studies addresses specifically the issue of discovering parallel pairs of articles in Wikipedia.

In this paper, we describe PARADOCS, a system capable of mining parallel documents in a collection, based on lightweight content-based features extracted from the documents. On the contrary to other systems designed to target parallel corpora (Chen

[1] http://www.statmt.org/wmt10

[2] http://opus.lingfil.uu.se/

[3] http://www.statmt.org/europarl/

[4] http://fr.wikipedia.org/

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 87–95,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

and Nie, 2000; Resnik and Smith, 2003), we do not assume any specific naming conventions on filenames or URLs.

The reminder of this article is organized as follows. In the next section, we describe our approach to mining parallel documents in a bilingual collection of texts. We test our approach on the `Europarl` corpus in section 3. We present in section 4 the application of our system to a subpart of the French-English articles of `Wikipedia`. We describe related work in section 5, summarize our work in section 6 and present future works in section 7.

2 PARADOCS

In order to identify pairs of parallel documents in a bilingual collection of texts, we designed a system, named PARADOCS, which is making as few assumptions as possible on the language pair being considered, while still making use of the content of the documents in the collection. Our system is built on three lightweight components. The first one searches for target documents that are more likely parallel to a given source document (section 2.1). The second component classifies (candidate) pairs of documents as parallel or not (section 2.2). The third component is designed to filter out some (wrongly) recognized parallel pairs, making use of collection-level information (section 2.3).

2.1 Searching Candidate Pairs

In a collection containing n documents in a given language, and m in another one, scoring each of the $n \times m$ potential pairs of source-target documents becomes rapidly intractable. In our approach, we resort to an information retrieval system in order to select the target documents that are most likely parallel to a given source one. In order to do so, we index target documents t in the collection thanks to an *indexing strategy* ϕ that will be described shortly. Then, for a source document s, we first index it, that is, we compute $\phi(s)$, and query the retrieval engine with $\phi(s)$, which in turn returns the N most similar target documents found in the collection. In our experiments, we used the `Lucene`[5] retrieval library.

[5] `http://lucene.apache.org`

We tested two indexing strategies: one reduces a document to the sequence of hapax words it contains ($\phi \equiv$`hap`), the other one reduces it to its sequence of numerical entities ($\phi \equiv$`num`). Hapax words have been found very useful in identifying parallel pairs of documents (Enright and Kondrak, 2007) as well as for word-aligning bitexts (Lardilleux and Lepage, 2007). Following Enright and Kondrak (2007), we define hapax words as blank separated strings of more than 4 characters that appear only once in the document being indexed. Also, we define a numerical entity as a blank separated form containing at least one digit. It is clear from this description that our indexing strategies can easily be applied to many different languages.

2.2 Identifying candidate pairs

Each candidate pair delivered by `Lucene`, is classified as parallel or not by a classifier trained in a supervised way to recognize parallel documents. Here again, we want our classifier to be as agnostic as possible to the pair of languages considered. This is why we adopted very light feature extractors ψ which are built on three types of entities in documents: numerical entities ($\psi \equiv$`num`), hapax words ($\psi \equiv$`hap`) and punctuation marks[6] ($\psi \equiv$`punc`). For each sequence of entities $\psi(s)$ and $\psi(t)$ of a source document s and a target document t respectively, we compute the three following features:

- the normalized edit-distance between the two representations:

  $$\sigma = ed(\psi(s), \psi(t))/max(|\psi(s)|, |\psi(t)|)$$

  where $|\psi(d)|$ stands for the size of the sequence of entities contained in d. Intuitively, σ gives the proportion of entities shared across documents,

- the total number of entities in the representation of both documents:

  $$|\psi(s)| + |\psi(t)|$$

  We thought this information might complement the one of σ which is relative to the document's sequence length.

[6] We only considered the 6 following punctuation marks that are often preserved in translation: `.!?():`

- A binary feature which fires whenever the pair of documents considered receives the smaller edit-distance among all the pairs of documents involving this source document:

$$\delta(s,t) = \begin{cases} 1 & \text{if } \mathrm{ed}\,(\psi(s),\psi(t)) \leq \mathrm{ed}\,(\psi(s),\psi(t'))\ \forall\, t' \\ 0 & \text{otherwise} \end{cases}$$

  Intuitively, the target document considered is more likely the good one if it has with the source document the smallest edit distance. Since we do compute edit-distance for all the candidate documents pairs, this feature comes at no extra computational cost.

We compute these three features for each sequence of entities considered. For instance, if we represent a document according to its sequence of numerical entities and its hapax words, we do compute a total of 6 features.[7]

It is fair to say that our feature extraction strategy is very light. In particular, it does not capitalize on an existing bilingual lexicon. Preliminary experiments with features making use of such a lexicon turned out to be less successful, due to issues in the coverage of the lexicon (Patry and Langlais, 2005).

To create and put to the test our classifier, we used the free software package `Weka` (Hall et al., 2009), written in Java.[8] This package allows the easy experimentation of numerous families of classifiers. We investigated logistic regression (`logit`), naive bayes models (`bayes`), adaboost (`ada`), as well as decision tree learning (`j48`).

2.3 Post-treatments

The classifiers we trained label each pair of documents independently of other candidate pairs. This independence assumption is obviously odd and leads to situations where several target documents are paired to a given source document and vice-versa. Several solutions can be applied; we considered two simple ones in this work. The first one, hereafter named `nop`, consists in doing nothing; therefore leaving potential duplicates source or target documents. The second solution, called `dup`, filters out pairs sharing documents. Another solution we did not implement would require to keep from the set of pairs concerning a given source document the one with the best score as computed by our classifier. We leave this as future work.

3 Controlled Experiments

We checked the good behavior of PARADOCS in a controlled experimental setting, using the `Europarl` corpus. This corpus is organized into bitexts, which means that we have a ground truth against which we can evaluate our system.

3.1 Corpus

We downloaded version 5 of the `Europarl` corpus.[9] Approximatively 6 000 documents are available in 11 languages (including English), that is, we have 6 000 bitexts in 10 language pairs where English is one of the languages. The average number of sentences per document is 273. Some documents contain problems (encoding problems, files ending unexpectedly, etc.). We did not try to cope with this. In order to measure how sensible our approach is to the size of the documents, we considered several slices of them (from 10 to 1000 sentences). [10]

3.2 Protocol

We tested several experimental conditions, varying the language pairs considered (en-da, -de, -el, -es, -fi, -fr, -it, -nl, -pt and -sv) as well as the document length (10, 20, 30, 50, 70, 100 and 1 000 sentences). We also tested several system configurations, varying the indexing strategy (`num`, `hap`), the entities used for representing documents (`hap`, `num`, `num+hap`, `num+punc`), the classifier used (`logit`, `ada`, `bayes`, and `j48`), as well as the post-filtering strategy (`nop`, `dup`). This means that we conducted no less than 4 480 experiments.

Because we know which documents are parallel, we can compute precision (percentage of identified parallel pairs that are truly parallel) and recall (percentage of true parallel pairs identified) for each configuration.

[7] We tried with less success to compute a single set of features from a representation considering all entities.

[8] `www.cs.waikato.ac.nz/ml/weka/`

[9] `http://www.statmt.org/europarl`

[10] We removed the first sentences of each document, since they may contain titles or other information that may artificially ease pairing.

Since our approach requires to train a classifier, we resorted in this experiment to a 5-fold cross-validation procedure where we trained our classifiers on 4/5 of the corpus and tested on the remaining part. The figures reported in the reminder of this section are averaged over the 5 folds. Also, all configurations tested in this section considered the $N = 20$ most similar target documents returned by the retrieval engine for each source document.

3.3 Results

3.3.1 Search errors

We first measured search errors observed during step 1 of our system. There are actually two types of errors: one when no document is returned by `Lucene` (`nodoc`) and one when none of the target documents returned by the retrieval engine are sanctioned ones (`nogood`). Figure 1 shows both error types for the Dutch-English language pair, as a function of the document length.[11] Clearly, search errors are more important when documents are short. Approximatively a tenth of the source documents of (at most) 100 sentences do not receive by `Lucene` any target document. For smaller documents, this happens for as much as a third of the documents. Also, it is interesting to note that in approximatively 6% of the cases where `Lucene` returns target documents, the good one is not present. Obviously we pay the prize of our lightweight indexation scheme. In order to increase the recall of our system, `nodoc` errors could be treated by employing an indexing strategy which would use more complex features, such as sufficiently rare words (possibly involving a keyword test, *e.g.* tf.idf). This is left as future work.

3.3.2 Best System configuration

In order to determine the factors which influence the most our system, we varied the language pairs (10 values) and the length of the documents (7 values) and counted the number of times a given system configuration obtained the best f-measure over the 70 tests we conducted. We observed that most of the time, the configurations recording the best f-measure are those that exploit numerical entities (both at indexing time and feature extraction time). Actually, we observed that computing features on

[11] Similar figures have been observed for other language pairs.

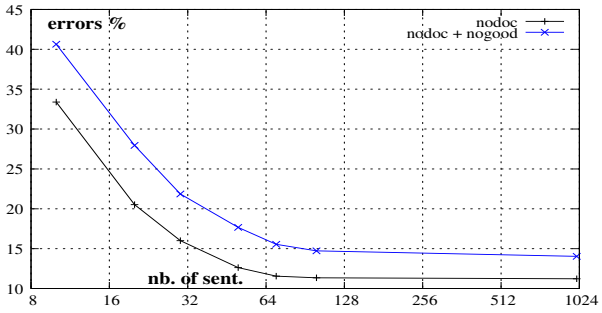

Figure 1: Percentage of Dutch documents for which `Lucene` returns no English document (`nodoc`), or no correct document (`nodoc+nogood`) as a function of the document size counted in sentences.

hapax words or punctuation marks on top of numerical entities do not help much. One possible explanation is that often, and especially within the `Europarl` corpus, hapax words correspond to numerical entities. Also, we noted that frequently, the wining configuration is the one embedding a logistic regression classifier, tightly followed by the decision tree learner.

3.3.3 Sensitivity to the language pair

We also tested the sensibility of our approach to the language pair being considered. Apart from the fact that the French-English pair was the easiest to deal with, we did not notice strong differences in performance among language pairs. For documents of at most 100 sentences, the worst f-measure (0.93) is observed for the Dutch/English language pair, while the best one (0.95) is observed for the French-English pair. Slightly larger differences were measured for short documents.

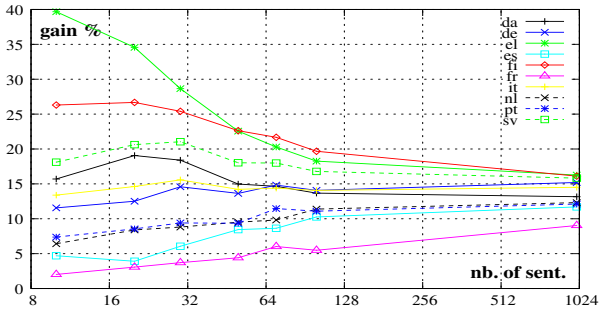

Figure 2: Absolute gains of the best variant of our system over the approach described by Enright and Kondrak (2007).

3.3.4 Sanity check

We conducted a last sanity check by comparing our approach to the one of (Enright and Kondrak, 2007). This approach simply ranks the candidate pairs in decreasing order of the number of hapax words they share. The absolute gains of our approach over theirs are reported in Figure 2, as a function of the document length and the language pair considered. Our system systematically outperforms the hapax approach of (Enright and Kondrak, 2007) regardless of the length of the documents and the language pairs considered. An average absolute gain of 13.6% in f-measure is observed for long documents, while much larger gains are observed for shorter ones. It has to be noted, that our approach requires to train a classifier, which makes it potentially less useful in some situations. Also, we used the best of our system in this comparison.

4 Experiments with Wikipedia

Many articles in `Wikipedia` are available in several languages. Often, they are explicitly marked as linked across languages. For instance, the English article *[Text_corpus]* is linked to the French one *[Corpus]*, but they are not translation of each other, while the English article *[Decline_of_the_Roman_Empire]* and the French one *[Déclin_de_l'empire_romain_d'Occident]* are parallel.[12]

4.1 Resource

During summer 2009, we collected all French-English cross-language linked articles from `Wikipedia`. A very straightforward preprocessing stage involving simple regular expressions removed part of the markup specific to this resource. We ended up with 537 067 articles in each language. The average length of the English pages is 711 words, while the average for French is 445 words. The difference in length among linked articles has been studied by Filatova (2009) on a small excerpt of bibliographical articles describing 48 persons listed in the biography generation task (Task 5) of DUC 2004.[13]

[12] At least they were at the time of redaction.

[13] `http://duc.nist.gov/duc2004/tasks.html/`

4.2 Parallelness of cross-language linked article pairs in FR-EN `Wikipedia`.

In this experiment, we wanted to measure the proportion of cross-language linked article pairs in `Wikipedia` that are in translation relation. In order to do so, we manually evaluated 200 pairs of articles in our French-English `Wikipedia` repository.

A web interface was developed in order to annotate each pair, following the distinction introduced by Fung and Cheung (2004): `parallel` indicates sentence-aligned texts that are in translation relation; `noisy` characterizes two documents that are nevertheless mostly bilingual translations of each other; `topic` corresponds to documents which share similar topics, but that are not translation of each others and `very-non` that stands for rather unrelated texts.

The results of the manual evaluation are reported in the left column of table 1. We observe that a fourth of the pairs of articles are indeed parallel or noisy parallel. This figure quantifies the observation made by Adafre and de Rijke (2006) that while some articles in `Wikipedia` tend to be translations of each other, the majority of the articles tend to be written independently of each other. To the best of our knowledge, this is the first time someone is measuring the degree of parallelness of `Wikipedia` at the article level.

If our sample is representative (something which deserves further investigations), it means that more than 134 000 pairs of documents in the French-English `Wikipedia` are parallel or noisy parallel.

We would like to stress that, while conducting the manual annotation, we frequently found difficult to label pairs of articles with the classes proposed by Fung and Cheung (2004). Often, we could spot a few sentences translated in pairs that we rated `very-non` or `topic`. Also, it was hard to be consistent over the annotation session with the distinction made between those two classes. Many articles are divided into sub-topics, some of which being covered in the other article, some being not.

4.3 Parallelness of the article pairs identified by PARADOCS

We applied PARADOCS to our `Wikipedia` collection. We indexed the French pages with the `Lucene`

	Wikipedia		PARADOCS	
Type	Count	Ratio	Count	Ratio
very-non	92	46%	5	2.5%
topic	58	29%	34	17%
noisy	22	11%	39	19.5%
parallel	28	14%	122	61%
Total	200		200	

Table 1: Manual analysis of 200 pairs cross-language linked in Wikipedia (left) and 200 pairs of articles judged parallel by our system (right).

toolkit using the num indexing scheme. Each English article was consequently transformed with the same strategy before querying Lucene, which was asked to return the $N = 5$ most similar French articles. We limited the retrieval to 5 documents in this experiment in order to reduce computation time. As a matter of fact, running our system on Wikipedia took 1.5 days of computation on 8 nodes of a pentium cluster. Most of this time was devoted to compute edit-distance features.

Each candidate pair of articles was then labeled as parallel or not by a classifier we trained to recognize parallel documents in an in-house collection of French-English documents we gathered in 2009 from a website dedicated to Olympic games.[14] Using a classifier trained on a different task gives us the opportunity to see how our system would do if used out-of-the-box. A set of 1844 pairs of documents have been automatically aligned (at the document level) thanks to heuristics on URL names; then manually checked for parallelness. The best classifier we developed on this collection (thanks to a 5-fold cross-validation procedure) was a decision tree classifier (j48) which achieves an average f-measure of 90% (92.7% precision, and 87.4% recall). This is the classifier we used in this experiment.

From the 537 067 English documents of our collection, 106 896 (20%) did not receive any answer from Lucene (nodoc). A total of 117 032 pairs of documents were judged by the classifier as parallel. The post-filtering stage (dup) eliminated slightly less than half of them, leaving us with a total of 61 897 pairs. We finally eliminated those pairs that were not cross-language linked in Wikipedia. We ended up with a set of 44 447 pairs of articles identified as parallel by our system.

Since there is no reference telling us which cross-language linked articles in Wikipedia are indeed parallel, we resorted to a manual inspection of a random excerpt of 200 pairs of articles identified as parallel by our system. The sampling was done in a way that reflects the distribution of the scores of the classifier over the pairs of articles identified as parallel by our system.

The results of this evaluation are reported in the right column of table 1. First, we observe that 20% (2.5+17) of the pairs identified as parallel by our system are at best topic aligned. One explanation for this is that topic aligned articles often share numbers (such as dates), sometimes in the same order, especially in bibliographies that are frequent in Wikipedia. Clearly, we are paying the prize of a lightweight content-oriented system. Second, we observe that 61% of the annotated pairs were indeed parallel, and that roughly 80% of them were parallel or noisy parallel. Although PARADOCS is not as accurate as it was on the Europarl corpus, it is still performing much better than random.

4.4 Further analysis

We scored the manually annotated cross-language linked pairs described in section 4.2 with our classifier. The cumulative distribution of the scores is reported in table 2. We observe that 64% (100-35.7%) of the parallel pairs are indeed rated as parallel ($p \geq 0.5$) by our classifier. This percentage is much lower for the other types of article pairs. On the contrary, for very non-parallel pairs, the classi-

	$p \leq 0.1$	$p \leq 0.2$	$p < 0.5$	avr.
very-non	1.1%	91.4%	92.5%	0.25
topic	1.7%	74.6%	78.0%	0.37
noisy	13.6%	77.3%	90.9%	0.26
parallel	7.1%	25.0%	35.7%	0.71

Table 2: Cumulative distribution and average score given by our classifier to the 200 manually annotated pairs of articles cross-language linked in Wikipedia.

[14] http://www.olympic.org

fier assigns a score lower than 0.2 in more than 91% of the cases. This shows that the score given by the classifier correlates to some extent with the degree of parallelness of the article pairs.

Among the 28 pairs of cross-language linked article pairs manually labelled as parallel (see table 1), only 2 pairs were found parallel by PARADOCS, even if 18 of them received a score of 1 by the classifier. This discrepancy is explained in part by the filter (`dup`) which is too drastic since it removes all the pairs sharing one document. We already discussed alternative strategies. The retrieval stage of our system is as well responsible of some failures, especially since we considered the 5 first French documents returned by `Lucene`. We further inspected the 10 (28-18) pairs judged parallel but scored by our classifier as non parallel. We observed several problems; the most frequent one being a failure of our pre-processing step which leaves undesired blocs of text in one of the article, but not in the other (recall we kept the preprocessing very agnostic to the specificities of `Wikipedia`). These blocs might be infoboxes or lists recapitulating important dates, or even sometimes HTML markup. The presence of numerical entities in those blocs is confounding the classifier.

5 Related Work

Pairing parallel documents in a bilingual collection of texts has been investigated by several authors. Most of the previous approaches for tackling this problem capitalize on naming conventions (on file URL names) for pairing documents. This is for instance the case of PTMINER (Chen and Nie, 2000) and STRAND (Resnik and Smith, 2003), two systems that are intended to mine parallel documents over the Web. Since heuristics on URL names does not ensure parallelness, other cues, such as the ratio of the length of the documents paired or their HTML structure, are further being used. Others have proposed to use features computed after sentence aligning a candidate pair of documents (Shi et al., 2006), a very time consuming strategy (that we tried without success). Others have tried to use bilingual lexicons in order to compare document pairs; this is for instance the case of the BITS system (Ma and Liberman, 1999). Also, Enright and Kondrak (2007) propose a very lightweight content-based approach to pairing documents, capitalizing on the number of hapax words they share. We show in this study, that this approach can easily be outperformed.

Zhao and Vogel (2002) were among the first to report experiments on harvesting comparable news collections in order to extract parallel sentences. With a similar goal, Munteanu et al. (2004) proposed to train in a supervised way (using some parallel data) a classifier designed to recognize parallel sentences. They applied their classifier on two monolingual news corpora in Arabic and English, covering similar periods, and showed that the parallel material extracted, when added to an in-domain parallel training corpus of United Nation texts, improved significantly an Arabic-to-English SMT system tested on news data. Still, they noted that the extracted material does not come close to the quality obtained by adding a small out-domain parallel corpus to the in-domain training material. Different variants of this approach have been tried afterwards, *e.g.* (Abdul-Rauf and Schwenk, 2009).

To the best of our knowledge, Adafre and de Rijke (2006) where the first to look at the problem of extracting parallel sentences from `Wikipedia`. They compared two approaches for doing so that both search for parallel sentence pairs in cross-language linked articles. The first one uses an MT engine in order to translate sentences of one document into the language of the other article; then parallel sentences are selected based on a monolingual similarity measure. The second approach represents each sentence of a pair of documents in a space of hyperlink anchored texts. An initial lexicon is collected from the title of the articles that are linked across languages (they also used the `Wikipedia`'s redirect feature to extend the lexicon with synonyms). This lexicon is used for representing sentences in both languages. Whenever the anchor text of two hyperlinks, one in a source sentence, and one in a target sentence is sanctioned by the lexicon, the ID of the lexicon entry is used to represent each hyperlink, thus making sentences across languages sharing some representation. They concluded that the latter approach returns fewer incorrect pairs than the MT based approach.

Smith et al. (2010) extended these previous lines of work in several directions. First, by training a global classifier which is able to capture the ten-

dency of parallel sentences to appear in chunks. Second, by applying it at large on Wikipedia. In their work, they extracted a large number of sentences identified as parallel from linked pairs of articles. They show that this extra materiel, when added to the training set, improves a state-of-the-art SMT system on out-domain test sets, especially when the in-domain training set is not very large.

The four aforementioned studies implement some heuristics in order to limit the extraction of parallel sentences to some fruitful document pairs. For news collections, the publication time can for instance be used for narrowing down the search; while for Wikipedia articles, the authors concentrate on document pairs that are linked across languages. PARADOCS could be used for narrowing the search space down to a set of parallel or closely parallel document pairs. We see several ways this could help the process of extracting parallel fragments. For one thing, we know that extracting parallel sentences from a parallel corpus is something we do well, while extracting parallel sentences from a comparable corpus is a much riskier enterprise (not even mentioning time issues). As a matter of fact, Munteanu et al. (2004) mentioned the inherent noise present in pairs of sentences extracted from comparable corpora as a reason why a large set of extracted sentence pairs does not contribute to improve an SMT system more that a small but highly specific parallel dataset. Therefore, a system like ours could be used to decide which sort of alignment technique should be used, given a pair of documents. For another thing, one could use our system to delimit a set of fruitful documents to harvest in the first place. The material acquired this way could then be used to train models that could be employed for extracting noisiest document pairs, hopefully for the sake of the quality of the material extracted.

6 Conclusion

We have described a system for identifying parallel documents in a bilingual collection. This system does not presume specific information, such as file (or URL) naming conventions, which can sometime be useful for mining parallel documents. Also, our system relies on a very lightweight set of content-based features (basically numerical entities and possibly hapax words), therefore our claim of a language neutral system.

We conducted a number of experiments on the Europarl corpus in order to control the impact of some of its hyper-parameters. We show that our approach outperforms the fair baseline described in (Enright and Kondrak, 2007). We also conducted experiments in extracting parallel documents in Wikipedia. We were satisfied by the fact that we used a classifier trained on another task in this experiment, but still got good results (a precision of 80% if we consider noisy parallel document pairs as acceptable). We conducted a manual evaluation of some cross-language linked article pairs and found that 25% of those pairs were indeed parallel or noisy parallel. This manually annotated data that can be downloaded at http://www.iro.umontreal.ca/~felipe/bucc11/.

7 Future Work

In their study on infobox arbitrage, Adar et al. (2009) noted that currently, cross-language links in Wikipedia are essentially made by volunteers, which explains why many such links are missing. Our approach lends itself to locate missing links in Wikipedia. Another extension of this line of work, admittedly more prospective, would be to detect recent vandalizations (modifications or extensions) operated on one language only of a parallel pair of documents.

Also, we think that there are other kinds of data on which our system could be invaluable. This is the reason why we refrained in this work to engineer features tailored for a specific data collection, such as Wikipedia. One application of our system we can think of, is the organization of (proprietary) translation memories. As a matter if fact, many companies do not organize the flow of the documents they handle in a systematic way and there is a need for tools able to spot texts that are in translation relation.

Acknowledgments

We are grateful to Fabienne Venant who participated in the manual annotation we conducted in this study.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 16–23.

Sisay Fissaha Adafre and Maarten de Rijke. 2006. Finding Similar Sentences across Multiple Languages in Wikipedia. In *11th EACL*, pages 62–69, Trento, Italy.

Eytan Adar, Michael Skinner, and Daniel S. Weld. 2009. Information arbitrage across multi-lingual wikipedia. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 94–103.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Jiang Chen and Jian-Yun Nie. 2000. Parallel Web text mining for cross-language IR. In *RIAO*, pages 62–67, Paris, France.

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

Tomaz Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *LREC*, Lisbon, Portugal.

Elena Filatova. 2009. Directions for exploiting asymmetries in multilingual wikipedia. In *Third International Cross Lingual Information Access Workshop*, pages 30–37, Boulder, Colorado.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 57–63, Barcelona, Spain, July. Association for Computational Linguistics.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11, Issue 1(10–18).

Philipp Koehn. 2005. Europarl: A multilingual corpus for evaluation of machine translation. In *10th Machine Translation Summit*, Phuket, Thailand, sep.

Adrien Lardilleux and Yves Lepage. 2007. The contribution of the notion of hapax legomena to word alignment. In *3rd Language & Technology Conference (LTC'07)*, pages 458–462, Poznań Poland.

Xiaoyi Ma and Mark Liberman. 1999. Bits: A method for bilingual text search over the web. In *Machine Translation Summit VII*, Singapore, sep.

Beata Bandmann Megyesi, Eva Csato Johansson, and Anna Sgvall Hein. 2006. Building a Swedish-Turkish Parallel Corpus. In *LREC*, Genoa, Italy.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May. Association for Computational Linguistics.

Alexandre Patry and Philippe Langlais. 2005. Automatic identification of parallel documents with light or without linguistic resources. In *18th Annual Conference on Artificial Intelligence (Canadian AI)*, pages 354–365, Victoria, British-Columbia, Canada.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380. Special Issue on the Web as a Corpus.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics (COLING) and the 44th annual meeting of the Association for Computational Linguistics (ACL)*, pages 489–496, Sydney, Australia.

Jason R. Smith, 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 NAACL*, HLT '10, pages 403–411.

Jörg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, pages 237–248. John Benjamins, Amsterdam/Philadelphia.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, ICDM '02, pages 745–748, Maebashi City, Japan.