Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora

Pierre Zweigenbaum[1] **Serge Sharoff**[2] **Reinhard Rapp**[3]
[1]LIMSI, CNRS, Université Paris-Saclay, Orsay, France
[2]University of Leeds, United Kingdom
[3]University of Mainz, Germany
pz@limsi.fr s.sharoff@leeds.ac.uk reinhardrapp@gmx.de

Abstract

In this paper we provide a summary of the rationale and the dataset contributing to the second shared task of the BUCC workshop. The shared task is aimed at detecting the best candidates for parallel sentences in a large text collection. The dataset for the shared task is based on a careful mix of parallel and non-parallel corpora. It contains 1.4 million French sentences and 1.9 million English sentences, in which 17 thousand sentence pairs are known to be parallel. The shared task itself is scheduled for the 2017 edition of the workshop.
Keywords: Parallel corpora, cross-language similarity

1. Introduction

Shared tasks gained importance in the NLP community since the data turn in the 1990s. They provide a way to compare different approaches using a common dataset and evaluation methods. In the field of comparable corpora, there are several options for shared tasks, such as:

1. methods for collecting comparable corpora from the Web;
2. methods for assessing the similarity of documents across languages in a collection of texts;
3. methods for assessing the similarity of separate sentences across languages in comparable corpora;
4. methods for detecting translations of words and phrases across languages in comparable corpora.

While it is difficult to operationalise the first task in this list, the 2015 edition of the BUCC workshop included the second task from this list (Sharoff et al., 2015). In it we used aligned Wikipedia articles to test document-level comparability methods for linking Chinese, French, German, Russian and Turkish articles to English.

In 2016 we aimed at building resources to test sentence-level comparability approaches. This paper describes our rationale for designing these resources, the methods used to build them, and the resulting data. A shared task based on these resources is planned for BUCC 2017.

2. Objectives

Our objectives were to create a dataset to evaluate parallel sentence extraction from comparable corpora.

Most former research on parallel sentence extraction from comparable corpora has relied on specific properties of the corpora used. This includes date properties in synchronous comparable corpora, e.g., international news in the same range of dates (Utiyama and Isahara, 2003; Munteanu et al., 2004; Abdul-Rauf and Schwenk, 2009), or document-level parallelism, e.g., encyclopedia articles for matched entries in two languages, as in Wikipedia.

The dependency on these specific properties creates two problems in our opinion. At a first level, we observe that these properties vary with the addressed corpora, and that they add to the difficulty of assessing the behavior of parallel sentence extraction methods. At a deeper level, we consider that the 'pure' task of translation spotting in comparable corpora should focus on content-based properties of the texts, not on external metadata.

The objective of the targeted task is therefore to test the ability of methods to detect parallel sentences in pairs of monolingual corpora *without using any metadata* on the corpora. In this task, only intrinsic properties of the sentences can be used.

Our initial design includes the following criteria, which we further refine and complement below after a study of related work:

- We start from two comparable corpora: these should not be the result of translations, as far as possible.
- No structural clues are provided beyond the order of sentences, which aims to be natural: the dataset provides no pre-existing document alignment (as in date-synchronized news or in linked Wikipedia pages).
- To be able to evaluate systems which detect parallel sentences in this pair of corpora, we need to know all 'positive examples' of parallel sentence pairs in these corpora. Therefore, we decided to introduce known pairs of parallel sentences into these comparable corpora.

3. Related work

This section briefly reviews related work relevant to the preparation of a corpus for the detection of parallel sentences.

3.1. Plagiarism detection: PAN

Shared tasks on plagiarism detection, as embodied by the PAN series (e.g., Potthast et al. (2012)), aim to detect instances of 'text re-use': text borrowed from one text into another. From the first editions on, PAN datasets have included not only monolingual but also cross-language instances of text re-use (Potthast et al., 2011).

The problem of detecting cross-language text re-use can be formulated as follows: does a text re-use parts of a previous text in a different language? It can be addressed as an 'intrinsic' cross-language plagiarism detection task,

Proceedings of the Ninth Workshop on Building and Using Comparable Corpora, pages 38–43,
LREC 2016, Portorož, Slovenia, May 23, 2016.

where 'translationese' is differentiated from original language (Barrón Cedeño, 2012, p. 145): methods for monolingual plagiarism detection can apply, such as differences in the distribution of function words or in language models. What Barrón Cedeño (2012, p. 147) names 'external' cross-language plagiarism detection is equivalent to the task of detecting text fragments with a high level of comparability (in particular parallel and highly comparable) from a multilingual corpus. In other words, we could consider that external cross-language text re-use and text alignment can be addressed as the same task, viewed from two different perspectives. Barrón Cedeño (2012) outlines five types of methods:

1. Models based on 'Syntax' (actually, morphology):
   Character dot-plot
   Character n-grams
   Cognateness
2. Models based on Thesauri (= single-word or term translation):
   EuroWordNet thesaurus
   Eurovoc thesaurus
3. Models based on Comparable Corpora (actually, aligned non-translated documents, namely Wikipedia)
   Cross-language explicit semantic analysis
4. Models based on Parallel Corpora:
   Bilingual representation space: Cross-language latent semantic indexing
   Bilingual mapping: Cross-language kernel canonical correlation analysis
5. Models based on Machine Translation (MT): Language normalisation (i.e., translation into one language)
   Web-based cross-language models (same as above, using on-line MT service)
   Multiple translations (i.e., output of MT before language model, with multiple translation hypotheses)

The present BUCC task is different from the PAN cross-language plagiarism detection in the following ways:

- The BUCC task aims to evaluate 'external' cross-language detection, whereas PAN is interested in both 'intrinsic' and 'external' cross-language plagiarism detection. As a consequence, the BUCC dataset should reduce the ease with which intrinsic plagiarism detection methods could spot artificially introduced sentences.
- The BUCC task focuses on sentence-level text fragments, whereas this granularity is not required by PAN.

3.2. Semantic text similarity: SemEval 2016

Semantic text similarity assesses the semantic equivalence of two texts or text fragments, e.g. sentences. Cross-language sentence similarity is close to evaluating whether two sentences are translations of one another: if they are, they obtain maximum similarity.

SemEval 2016[1] includes a cross-language sentence similarity task: its goal is to evaluate the similarity of sentence pairs which belong to two different languages, instantiated on the English-Spanish language pair. The task is formulated as scoring a given pair of sentences on a six-point scale.

The trial data was drawn from sentence pairs used in prior English semantic text similarity evaluations (STS 2012, 2013, 2014, and 2015). Bilingual data was obtained by translating some of the English sentences into Spanish and considering that the semantic similarity score for a resulting cross-lingual pair was that of the original English sentence pair, then filtering out some lower quality cross-lingual pairs.

Note that the interpretation of the scores can be related to comparability. The examples provided by the organizers and their explanations of the scores are copied below, from highest to lowest similarity *(We added an English translation of the Spanish sentences in parentheses.)*. The highest similarity score (5) corresponds to an exact translation, the next (4) is probably an acceptable translation, whereas the following one (3) would be an inexact translation and would be likely to obtain a lower BLEU score. Sentence pairs with lower scores would be likely to introduce too much noise if added to the training corpus of a statistical machine translation system.

(5) The two sentences are completely equivalent, as they mean the same thing
El pájaro se está bañando en el lavabo. *(The bird is washing itself in the water basin.)*
Birdie is washing itself in the water basin.

(4) The two sentences are mostly equivalent, but some unimportant details differ.
En mayo de 2010, las tropas intentaron invadir Kabul. *(In May 2010, the troops tried to invade Kabul.)*
The US army invaded Kabul on May 7th last year, 2010.

(3) The two sentences are roughly equivalent, but some important information differs or is missing.
John dijo que él es considerado como testigo, y no como sospechoso. *(John said that he is considered as a witness, not as a suspect..)*
"He is not a suspect anymore." John said.

(2) The two sentences are not equivalent, but share some details.
Ellos volaron del nido en grupos. *(They flew from the nest in groups.)*
They flew into the nest together.

(1) The two sentences are not equivalent, but are on the same topic.
La mujer está tocando el violín. *(The woman is playing the violin.)*
The young lady enjoys listening to the guitar.

(0) The two sentences are on different topics.
Al amanecer, Juan se fue a montar a caballo con un grupo de amigos. *(At dawn, Juan went riding with a group of friends.)*
Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

[1] http://alt.qcri.org/semeval2016/task1/

The README of the task suggests methods to compute cross-language similarity: Adapting monolingual 'align+featurize' semantic text similarity systems to the cross-lingual task; deep learning with cross-lingual embeddings; and monolingual semantic text similarity complemented with machine translation.
The present BUCC task has two differences with cross-language semantic text similarity:

- The BUCC task uses a binary scale to evaluate whether or not two sentences are translations of each other.
- The BUCC task does not provide a list of sentence pairs, but instead provides two monolingual lists of sentences. The set of sentence pairs to be examined by the systems is potentially the cross-product of these two sets of sentences: this creates the need for efficient comparison or pruning methods.

3.3. Bilingual document alignment: WMT 2016

WMT 2016 includes a shared task on bilingual document alignment[2]. In that task, given two sets of Web pages in two languages from the same Web domain, each pair of translated source-target page pairs must be detected.
Whereas the similarity to the BUCC task is clear, two differences can be noted:

- The main difference is the granularity of the documents to be aligned: the BUCC task addresses sentences, whereas WMT 2016 addresses documents (Web pages).
- Another difference however is that the BUCC task aims not to use any metadata; in contrast, WMT provides metadata on its documents: the Web page URLs, which make it possible to use non-content-based methods to address the task. As a matter of fact, an implementation of such a method is provided as a baseline by the organizers and can be downloaded from the WMT Web site.

4. Data preparation methods

An ecologically sound way to produce resources for our task would be by annotating manually parallel sentences in a large selection of sentences from a real comparable corpus, i.e., two comparable monolingual corpora. However, exact translations are very rare in a randomly collected corpus, and manually spotting them would be labor-intensive. Often they imply that their two provenant documents are reasonable translations (in either direction). To increase the probability of finding parallel sentences, the two corpora could thus be selected so that they consist of pairs of matching documents on the same topic. But many sentences in a collection of aligned Web documents are likely to originate from machine translated texts (Antonova and Misyurev, 2011). Additionally, detecting automatically translated sentences is easy if using the same MT system (primarily, Google Translate) (Potthast et al., 2012).

Therefore, we switched to creating a dataset which is prepared automatically from a known parallel corpus and known non-parallel sentences from two monolingual corpora. In this dataset, known pairs of parallel sentences are 'planted' into existing monolingual corpora.
The above-mentioned work on cross-lingual plagiarism suggests to invest some effort into a reasonably good blend of the planted sentences in their environment (what the plagiarism literature calls 'obfuscation'). Otherwise, it could be easier to identify which (parallel) sentences were added to the initial texts than to check their parallelism. To determine in which document a passage of another document can be inserted, Asghari et al. (2015) perform sentence and document clustering based on the sentence similarity obtained through Information Retrieval queries with the Lucene IR engine. We followed a similar though simpler approach to determine where to insert parallel sentences, which we describe below.
1. Indexing collections of monolingual sentences.

- Our initial data is composed of a pair of comparable monolingual corpora (Wikipedia dumps in two languages, say EN and FR) and a sentence-aligned parallel corpus in the same pairs of languages (News Commentary[3]).
- We split each of the two monolingual corpora into sentences (using the Europarl sentence splitter).
- We treated each monolingual corpus as a collection of sentences and indexed them with an information retrieval engine (Apache SolR[4]).

2. Spotting similar sentences through IR queries.

- For each pair of parallel sentences, we used the EN sentence as a query to the EN collection of sentences and the FR sentence as a query to the FR collection of sentences. If successful, this should identify a locations in the EN (resp. FR) monolingual corpus where the EN (resp. FR) parallel sentence could be inserted in a context where they have chances to be related to the current topic.
- Our motivation for using an IR engine is to implement a scalable sentence similarity computation process with minimal investment. We chose query parameters which impose stronger similarity constraints on the query (parallel sentence) and the 'document' (monolingual sentence), for instance by setting a minimum number of common content words (5) between query and document and imposing a similar total number of content words in both sentences.
- We consider a pair of queries as successful if it retrieves at least one EN sentence and one FR sentence with the chosen constraints.

3. Inserting parallel and non-parallel sentences into the monolingual corpora.

[2] http://www.statmt.org/wmt16/bilingual-task.html

[3] http://www.casmacat.eu/corpus/news-commentary.html

[4] http://lucene.apache.org/solr/

- Given a pair of locations (similar monolingual sentences) identified by a successful pair of queries built from a pair of aligned EN-FR sentences, we insert the parallel EN sentence after the monolingual EN sentence similar to it, and the parallel FR sentence before the monolingual FR sentence similar to it.

- After the previous step, each sentence inserted into one of the two monolingual corpora is parallel to a sentence inserted in the other monolingual corpus. This means that if a system detects inserted sentences (for instance through intrinsic plagiarism detection methods as mentioned in Section 3.1.), it can be certain that this sentence belongs to the gold standard. To reduce this certainty, we also insert A adjacent sentences from the parallel corpora together with the parallel EN and FR sentences: A sentences following the EN sentence and A sentences preceding the FR sentence, so that these added sentences are not parallel. Now not all inserted sentences are parallel anymore.[5]

4. Increasing the rate of parallel sentences.

   - In the above-described process, only a small proportion of sentences in the original comparable corpora become an insertion point for parallel sentences. To increase the rate of parallel sentences in the resulting corpus, we only keep monolingual documents (Wikipedia pages) where at least one insertion point has been found.

   - When a monolingual document is included in the corpus, if there is an interlanguage link from it to a document (Wikipedia page) in the other language, it is inserted too, even though no parallel sentence may have been inserted into that linked document.

   - Some monolingual documents are much longer than the others: to reduce the non-parallel part of the corpus further, we truncate them to their first 500 sentences.

5. Reducing the rate of non-inserted parallel sentences.

   - There is always a chance that naturally-occurring parallel sentences exist in a pair of Wikipedia pages. We need to know about them to be able to provide a fair evaluation of translation spotting systems. However detecting them automatically is the very goal of our target shared task, so we cannot assume we have a system which will do this perfectly. We envision several methods to reduce these pairs of naturally-occurring parallel sentences.

     1. Use an existing system to spot them and either add them to the gold standard or remove them from the data. A problem is that this will bias the corpus towards this system.
     2. Use an existing method or system with relaxed constraints to increase the recall of the detection of potentially parallel sentences, for instance by translating source sentences automatically to the target language and using a semantic text similarity metric (see Section 3.2.) to spot (and remove) pairs of sentences with a similarity above some relatively low similarity threshold (e.g., between 2 and 3 on the SemEval scale presented in Section 3.2.). A problem is that this will bias the distribution of cross-lingual sentence similarity, creating a gap between unrelated sentences and (inserted) translated sentences.
     3. At evaluation time, pool the results of the participating systems and have humans examine false-positive sentence pairs found by a consensus of at least N systems. This requires a human investment which remains to be estimated.

   - Since pairs of shorter sentences are more likely to be chance translations of each other, we removed from the corpus sentences with less than a ceiling of C content words.

5. Dataset

We instantiated the above-mentioned method on the French-English pair of languages:

- The monolingual corpora are July-August 2014 XML Wikipedia dumps provided by the LinguaTools Web site[6]. We prepared the text versions of these corpora by using the associated tool xml2txt[7]. HTML entities were converted into their UTF-8 equivalent. Documents were further tokenized[8] and split into sentences as detailed above. The English corpus contains 4.5M articles and 138M sentences, the French corpus 1.5M articles and 46M sentences.

- The parallel corpus comes from the News Commentary, version 9, provided as training data for WMT 2014[9]. The French-English News Commentary corpus contains 183k sentence pairs.

- After some experiments, we set the following Solr query parameters: efType="edismax", qs=5, ps=5, ps2=5, mm="70%", stopwords="true". With these parameters, the process retrieved similar sentences for 18k sentence pairs, representing 10% of the News Commentary sentence pairs and 0.03% of the French Wikipedia sentences.

- After completion of the process, the produced comparable corpora contain respectively 1.4M French sentences and 1.9M English sentences, including 17k inserted parallel sentences in each corpus.

[5] Indeed a system using intrinsic plagiarism detection methods might probably still spot the inserted passages and reduce the complexity of the search for parallel sentences. Again, this is not what the present shared task aims to evaluate.

[6] http://linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

[7] xml2txt.pl -articles -nomath -notables -nodisambig

[8] Tokenization is performed by the Solr indexer anyway and was not really necessary at this step.

[9] http://www.statmt.org/wmt14/training-parallel-nc-v9.tgz

French monolingual sentences	English monolingual sentences
fr-000000197 Si bien que l'année suivante, elle mit sa priorité dans les initiatives régionales telles que le Mercosur ou la Banque du Sud après une décennie de partenariat avec les États-Unis.	**en-001425664** All the while, scant attention is paid to the region's already established bodies, which are in sad shape.
fr-000000198 Prenons l'exemple du MERCOSUR (le Marché commun du Sud), la principale initiative régionale d'après-guerre.	**en-001425665** Consider MERCOSUR, the main post-Cold War regional initiative.
fr-000000199 Selon l'universitaire argentin Roberto Bouzaq, le MERCOSUR est dans un état critique en raison de son incapacité à maintenir le cap sur les objectifs communs qui ont conduit les pays-membres à s'engager dans un processus d'intégration régionale, avec pour conséquence un éparpillement et l'impossibilité d'identifier les problèmes politiques sous-jacents qui devraient être prioritaires.	**en-001876436** Indeed, this vision of international relations clearly rests on building influence through military power.
. . .	**en-001876437** Finally, Argentina stands in a category by itself.
fr-000000203 Enfin, l'Argentine fut l'un des signataires initiaux du Traité sur l'Antarctique.	
fr-000000204 Enfin, l'Argentine est un cas à part.	

Table 1: Example sentences: comparable corpora with inserted pairs of parallel sentences (see Table 2).

Table 1 shows an excerpt of our collection. Two out of four sentences in each language are linked in the gold alignment file, as shown in Table 2.

fr-000000198	⇔	en-001425665
fr-000000204	⇔	en-001876437

Table 2: Example gold standard alignments (sentence pairs from parallel corpus).

6. Limitations

The current design and its realization have the following limitations.

- The insertion of the parallel sentence pairs (from News Commentary) into the monolingual corpora (from Wikipedia) is sometimes coherent, sometimes not really coherent.
- At some point in the implementation of the method the monolingual corpora were tokenized, but not the parallel corpora. This created surface differences which can reveal the origin of sentences. 'Detokenizing' (pasting back punctuation to the adjacent token) is not an easy process, and we should reprocess the corpus without tokenization, which was not really needed in our pipeline.
- The need for obfuscating the inserted parallel sentence pairs remains a matter of debate. A much higher quality would be required of the blending of inserted sentences into the monolingual corpora than what was performed here, for instance as in (Asghari et al., 2015), for it to be really useful.
- Translation pairs that may exist naturally in Wikipedia are not removed nor known exactly, and may hence lead to counting false positives in the evaluation if systems find them. Human review of pooled system results are a possible solution to this problem, but require manpower.
- The above-described method was applied to the French-English language pair as a proof of concept. It is yet to be applied to other language pairs. This would be feasible in principle for German, Russian, and Chinese, for which source data are available in both Wikipedia and News Commentary (and which Solr can handle). Turkish is handled by Solr but is not present in the News Commentary collection of parallel corpora.

7. Evaluation method

The primary evaluation measure is the F-score of sentence pairs:

- A sentence pair is considered correct if it is present in the gold standard.
- Precision is the proportion of correct system-generated pairs among those pairs returned by the system.
- Recall is the proportion of correct system-generated pairs among all pairs in the gold standard.
- F is the harmonic mean of precision and recall.

8. Shared task plans

Because of the complexities involved in preparation of the dataset, the task initially proposed for the 2016 edition of the BUCC workshop had to be postponed to 2017.

References

Abdul-Rauf, S. and Schwenk, H. (2009). Exploiting comparable corpora with TER and TERp. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: From Parallel to Non-parallel Corpora*, BUCC '09, pages 46–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Antonova, A. and Misyurev, A. (2011). Building a web-based parallel corpus and filtering out machine-translated text. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 136–144.

Asghari, H., Khoshnava, K., Fatemi, O., and Faili, H. (2015). Developing bilingual plagiarism detection corpus using sentence aligned parallel corpus. In *Notebook for PAN at CLEF 2015*, Toulouse, France.

Barrón Cedeño, L. A. (2012). *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism*. Ph.D. thesis, Universitat Politècnica de València.

Munteanu, D. S., Fraser, A., and Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In Daniel Marcu Susan Dumais et al., editors, *HLT-NAACL 2004: Main Proceedings*, pages 265–272, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Potthast, M., Barrón-Cedeño, A., Stein, B., and Rosso, P. (2011). Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.

Potthast, M., Gollub, T., Hagen, M., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., and Stein, B. (2012). Overview of the 4th international competition on plagiarism detection. In Pamela Forner, et al., editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). BUCC shared task: Cross-language document similarity. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 74–78, Beijing, China, July. Association for Computational Linguistics.

Utiyama, M. and Isahara, H. (2003). Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 72–79, Sapporo, Japan, July. Association for Computational Linguistics.