Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora

Pierre Zweigenbaum
LIMSI, CNRS,
Université Paris-Saclay,
F-91405 Orsay, France
pz@limsi.fr

Serge Sharoff
University of Leeds,
Leeds, United Kingdom
s.sharoff@leeds.ac.uk

Reinhard Rapp
Magdeburg-Stendal University
of Applied Sciences and
University of Mainz, Germany
reinhardrapp@gmx.de

Abstract

This paper presents the BUCC 2017 shared task on parallel sentence extraction from comparable corpora. It recalls the design of the datasets, presents their final construction and statistics and the methods used to evaluate system results. 13 runs were submitted to the shared task by 4 teams, covering three of the four proposed language pairs: French-English (7 runs), German-English (3 runs), and Chinese-English (3 runs). The best F-scores as measured against the gold standard were 0.84 (German-English), 0.80 (French-English), and 0.43 (Chinese-English). Because of the design of the dataset, in which not all gold parallel sentence pairs are known, these are only minimum values. We examined manually a small sample of the false negative sentence pairs for the most precise French-English runs and estimated the number of parallel sentence pairs not yet in the provided gold standard. Adding them to the gold standard leads to revised estimates for the French-English F-scores of at most +1.5pt. This suggests that the BUCC 2017 datasets provide a reasonable approximate evaluation of the parallel sentence spotting task.

1 Introduction

Shared tasks and the associated datasets have proved their worth as a driving force in a number of subfields of Natural Language Processing. However, very few shared tasks were organized on the topic of comparable corpora. Therefore, we endeavored to design and organize shared tasks as companions of the BUCC workshop series on Building and Using Comparable Corpora. The First BUCC Shared Task (Sharoff et al., 2015) tackled the detection of comparable documents across languages. The Second BUCC Shared Task,[1] presented here, addresses the detection of parallel sentences across languages in non-aligned, monolingual corpora.

Let us recall the overall goals, design and principles of this task, which were introduced in (Zweigenbaum et al., 2016). A bottleneck in statistical machine translation is the scarceness of parallel resources for many language pairs and domains. Previous research has shown that this bottleneck can be reduced by utilizing parallel portions found within comparable corpora (Utiyama and Isahara, 2003; Munteanu et al., 2004; Abdul-Rauf and Schwenk, 2009). These are useful for many purposes, including automatic terminology extraction and the training of statistical MT systems. However, past work relied on meta-information, such as the publication date of news articles or inter-language links in Wikipedia documents, to help select promising sentence pairs before examining them more thoroughly. It is therefore difficult to separate the heuristic part of the methods that deals with this meta-information in clever ways from the cross-language part of the methods that deals with translation and comparability issues. We consider that the latter type of methods is more fundamental and wanted to focus on its evaluation. We thus designed a task in which no meta-information is available on the relation between the two monolingual corpora in which pairs of translated sentences are to be found.

In (Zweigenbaum et al., 2016) we showed the difference of this task to PAN's cross-language plagiarism detection (Potthast et al., 2012), SemEval's cross-language semantic text similarity

[1] https://comparable.limsi.fr/bucc2017/bucc2017-task.html

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 60–67,
Vancouver, Canada, August 3, 2017.

(Agirre et al., 2016), and WMT's bilingual document alignment (Buck and Koehn, 2016).

The present paper reports the actual organization of the task as a companion to the BUCC 2017 workshop. We describe the final method we used to prepare bilingual corpora in four language pairs: Chinese-English, French-English, German-English, and Russian-English (Section 2), the evaluation method (Section 3), the participants' systems (Section 4), the results they obtained (Section 5), and conclude (Section 6).

2 Corpus preparation

The challenges we faced to prepare corpora for a parallel sentence spotting shared task, and the measures we took to address them, were the following.

1. Given two monolingual corpora, it would be very long for human evaluators to find all sentence pairs that are translations of each other. Therefore we decided to insert known parallel sentence pairs into existing monolingual corpora. We chose Wikipedia articles (20161201 dumps [2]) as our monolingual corpora and News Commentary (v11[3]) as our source for parallel sentence pairs. In the remainder of this section we use French and English as a running example of a language pair.

2. These inserted parallel sentence pairs should not be trivially detectable in the monolingual corpora. Therefore we strove to insert sentences that are coherent with the context in which they are inserted. In this purpose we aimed to select as insertion points sentences that were similar in topic to the inserted sentences. We implemented this by indexing with the Solr search engine each English sentence of the monolingual corpus (English Wikipedia dump, converted to text and split into sentences) and each French sentence of the monolingual corpus (French Wikipedia dump, converted to text and split into sentences). For each sentence pair in the parallel corpus (French-English News Commentary), we queried Solr to find the most similar French sentence and English sentence for this pair; if hits were found for both languages, we recorded as insertion point for the French parallel sentence the French sentence found, and as insertion point for the English parallel sentence the English sentence found. We performed the actual insertion after all parallel sentence pairs were thus processed.

Additionally, a different distribution of sentence lengths in the original monolingual sentences and in the inserted sentences might give hints about the origin of a sentence. Therefore we aimed at having similar distributions of sentence lengths for both the Wikipedia sentences and the News Commentary sentences. In this purpose, we excluded sentences outside a range of lengths (we kept sentences between 20 and 40 words long).

We also tried to reduce trivial typographical differences that may be revealing of the source of a sentence, such as the use of certain quotation marks and certain systematic conversion issues found in Wikipedia texts after conversion from their Wiki source. In this purpose we customized an existing Wikipedia conversion tool, WikiExtractor.py,[4] to include sentence splitting (with NLTK). Since template processing was the cause of a large number of idiosyncrasies in the converted Wikipedia text, we removed the sentences that contained a template.

3. The original monolingual texts should contain as few 'natural' parallel sentence pairs as possible. Since interlinked Wikipedia articles are a common source of parallel sentence pairs, we ensured that a given dataset never contained sentences from such a pair of documents.

4. When the two sentences in a parallel pair are inserted in the monolingual corpora, there is no particular reason for them to be positioned in similar locations in the two corpora. Therefore, once a corpus has been generated this way, splitting it into training and test would be likely to separate a number of parallel pairs. Besides, an additional small *sample* split was also needed for prospective participants to examine data and decide whether they would be interested, extending the problem further.

To prevent this problem, we split each pair of monolingual corpora, before indexing and parallel sentence insertion, into sample, training and test corpus pairs, respectively with 2%, 49% and 49% of the full corpora (the number and sizes of these splits are specified as parameters to the algorithm). Given as input two sets of Wikipedia pages, the algorithm randomly distributes them into the N splits according to the specified probabilities. It

[2] http://ftp.acc.umu.se/mirror/wikimedia.org/dumps/

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

[4] https://github.com/attardi/wikiextractor

Pair	Sample (2%)			Training (49%)			Test (49%)		
	fr	en	gold	*fr*	en	gold	*fr*	en	gold
de-en	32593	40354	1038	413869	399337	9580	413884	396534	9550
fr-en	21497	38069	929	271874	369810	9086	276833	373459	9043
ru-en	45459	72766	2374	460853	558401	14435	457327	566356	14330
zh-en	8624	13589	257	94637	88860	1899	91824	90037	1896

Table 1: Corpus statistics: number of monolingual sentences (*fr*, en) and of parallel pairs (gold) for each split and each language pair. The *fr* column stands for the non-English language in each pair.

Name	Affiliation	Language pairs
VIC	Vicomtech-IK4, Donostia / San Sebastian, Gipuzkoa, Spain	de-en (3), fr-en (3)
RALI	RALI - DIRO, Université de Montréal, Montréal, Québec, Canada	fr-en (3)
JUNLP	Department of Computer Science and Engineering, Jadavpur University, India	fr-en (1)
zNLP	LIMSI, CNRS, Université Paris-Saclay, Orsay, France	zh-en (3)

Table 2: Shared task systems

also ensures that no interlinked pair of pages is distributed to the same split. Indexing, searching and sentence insertion were then performed on each split separately. Since the training and test sets for a given language pair were generated with the same process and parameters, they received very similar numbers of parallel sentence pairs.

This process was applied to five languages (Chinese (zh), English (en), French (fr), German (de), Russian (ru)) to produce four bilingual datasets, each split into sample, training, and test data. Table 1 shows the statistics of the resulting datasets.

3 Evaluation method

Given two sentence-split monolingual corpora, participant systems were expected to identify pairs of sentences that are translations of each other. Each team was allowed to submit up to three runs per language pair.

Evaluation was performed using balanced F-score. In the results of a system, a true positive TP is a pair of sentences that is present in the gold standard and a false positive FP is a pair of sentences that is not present in the gold standard. A false negative FN is a pair of sentences present in the gold standard but absent from system results. Precision, Recall and F1-score were then computed using the usual formulas.

Of note, this evaluation is performed on the synthetic corpus presented above, using the inserted parallel sentence pairs as the gold standard. Therefore it does not take into account the possible existence of true parallel pairs present in the monolingual corpora beyond the inserted sentence pairs. By avoiding aligned Wikipedia articles, the construction of the corpus attempted to reduce the likelihood of such sentence pairs, but indeed it did not suppress it altogether. For these reasons we also performed a limited experiment in which human judges evaluated selected samples of the system results. The assessment of each sentence pair was performed according to the guidelines of the SemEval 2016 cross-language sentence similarity task (Agirre et al., 2016).

4 Participants and systems

About 17 teams downloaded datasets, among which four teams submitted runs: VIC (Spain) (Azpeitia et al., 2017), RALI (Canada) (Grégoire and Langlais, 2017), JUNLP (India) (Mahata et al., 2017), and LIMSI (France: 'zNLP') (Zhang and Zweigenbaum, 2017). Table 2 gives more detail about teams and runs.

All systems had to include a way to cope with the bilingual dimension of the task. This was addressed with pre-existing dictionaries (LIMSI), machine translation systems (JUNLP, LIMSI), word alignments obtained from parallel corpora (VIC), or bilingual word embeddings trained from parallel corpora (RALI).

Cross-language sentence similarity was then handled by Cosine similarity (JUNLP, LIMSI, RALI) or the Jaccard coefficient (VIC), possibly with weighting (a function of frequency: VIC; tf.idf: LIMSI) and with a trained classifier (RALI, LIMSI). Some teams used an Information Retrieval engine to accelerate the search for similar

sentences (VIC, LIMSI).

JUNLP (Mahata et al., 2017) implemented a baseline method that translates the FR corpus with a Machine Translation system, selects candidate sentence pairs with a suitable length ratio, and chooses the final sentence pairs based on Cosine similarity.

zNLP (Zhang and Zweigenbaum, 2017) used a bilingual dictionary to perform word-level translation of the ZH corpus, complemented by calls to an on-line Machine Translation system. They used the Solr search engine to index sentences and search for similar sentences, collecting a number of candidate translations for each 'source' sentence. They selected the best translation (or none) by training a classifier with Solr score and rank, word overlap, and sentence length features.

RALI (Grégoire and Langlais, 2017) experimented with a deep learning framework. They trained bilingual word embeddings with BilBOWA (Bilingual Bag-of-Words without Alignments (Gouws et al., 2015)) on the Europarl parallel corpus, represented source and target sentences in this common space and used Cosine similarity to select candidate parallel sentence pairs. They also trained a bidirectional recurrent neural network with gated recurrent units (BiGRU) on both the source and target languages to build sentence-level continuous representations. They learned a linear transformation of these representations from one language to the other and decided on the parallelism of two sentences based on the comparison of their continuous representations through this transformation.

VIC (Azpeitia et al., 2017) used probabilistic dictionaries acquired by word alignment of parallel corpora to translate each corpus. They used the Lucune search engine to index sentences and search for similar sentences, collecting a number of candidate translations for each 'source' sentence, in both directions. Final sentence similarity is computed by their STACC method (Set-Theoretic Alignment for Comparable Corpora, (Etchegoyhen and Azpeitia, 2016)), which extends basic word overlap by taking into account non-matched words that share a long enough common prefix, as well as numbers and capitalized true-cased tokens. STACC measures word overlap with the Jaccard coefficient. They refined the STACC method by taking into account lexical weights that penalize frequent words.

5 Results and discussion

We first present an evaluation based upon the inserted translation pairs (Section 5.1) then an additional evaluation based upon human judgment of sample system results (Section 5.2)

5.1 Automatic evaluation

We present here the evaluation results for the submitted runs for each language in turn. As explained above, these results are based on the artificially inserted translation pairs. In each table we show the precision, recall and F1-score of each run in percentages. Because this synthetic dataset represents an approximation of a real task, there is no point in computing precise scores: we round the computed percentages to the nearest integer.

Additionally, we observed that some participants took into account the prior probability of translation pairs in the training datasets. Given that the test dataset was announced to be generated in the same way as the training dataset, they targeted a number of translation pairs in the test that was consistent with this prior probability. We therefore display this number of translation pairs in the tables too.

Three teams submitted runs on the French-English (fr-en) language pair. In addition to these runs, Table 3 presents the minimum, maximum, median, mean and standard deviation for each measure. The initial JUNLP submission had a bug which was fixed a couple of days later; we show the results of the fixed submission in italics, but did not include it in the additional statistics. The VIC results confirm the strategy described in

run_name	sys_n	P (%)	R (%)	F1 (%)
VIC1	8831	80	79	**79**
VIC2	7569	**87**	73	**79**
VIC3	10768	70	**83**	76
RALI2	47576	12	63	20
RALI1	57761	10	66	18
RALI3	66201	9	63	15
JUNLP1	38736	3	11	4
min	7569	9	63	15
median	29172	41	70	48
mean	33118	45	71	48
max	66201	87	83	79
stddev	24062	34	7	30

Table 3: Evaluation of fr-en runs (n_gold=9,043)

(Azpeitia et al., 2017) by which they optimized

VIC1 for F1-score, VIC2 for precision, and VIC3 for recall; the results for German also display the same pattern. The three runs RALI2, RALI1 and RALI3 produce an increasing number of candidate pairs, resulting in a decrease in precision; this leads to an increase in recall only for RALI1, but always to a decrease in F1-score. Reasons for the lower precisions and (to a lesser extent) recalls of the RALI results are proposed in (Grégoire and Langlais, 2017), including the handling of numbers (improved in their later experiments) and the selection of negative training examples.

Only one team submitted runs on the German-English (de-en) language pair, therefore we do not report min, max and other statistics. The results are displayed in Table 4. The precisions and

run_name	sys_n	P (%)	R (%)	F1 (%)
VIC1.de-en	8640	88	80	**84**
VIC3.de-en	9949	82	**85**	**84**
VIC2.de-en	7586	**92**	73	82

Table 4: Evaluation of de-en runs (n_gold=9,550)

F1-scores obtained by VIC for German-English are higher than those they obtained for French-English, with similar recalls. The only difference in the two corpora in terms of statistics is that the German-English dataset was more balanced in its numbers of monolingual sentences, but other differences linked to the intrinsic properties of German and French or to the resources used to train the system for these two languages are likely to have an effect too.

One team submitted runs on the Chinese-English (zh-en) language pair, therefore we do not report min, max and other statistics. The results are displayed in Table 5. According to (Zhang and

run_name	sys_n	P (%)	R (%)	F1 (%)
zNLP1	1985	42	**44**	**43**
zNLP3	1526	**46**	37	41
zNLP2	1900	19	19	19

Table 5: Evaluation of zh-en runs (n_gold=1,896)

Zweigenbaum, 2017), zNLP3 was optimized for precision: this is confirmed by its results on the test set. Overall, the results are lower than the best runs on the fr-en and de-en datasets. Various hypotheses can be proposed to account for this difference, including the different types and sizes of the resources used for translation in VIC and zNLP, the specific methods used in the two systems, and differences in intrinsic language properties.

5.2 Complementary human evaluation

Were we to know which 'natural' translation pairs existed in the test datasets beyond the translation pairs we inserted, would the results be very different? We did not have resources to perform an extensive human evaluation to answer this question, therefore we designed a minimal experiment on the French-English language pair.

In the VIC and RALI runs, we selected the run with the best precision and randomly drew 20-pair samples. A French native speaker with good command of English examined each sample and scored it according to the grades used in the SemEval 2016 cross-language sentence similarity task (Agirre et al., 2016): (5) The two sentences are completely equivalent, as they mean the same thing; (4) The two sentences are mostly equivalent, but some unimportant details differ; (3) The two sentences are roughly equivalent, but some important information differs or is missing; (2) The two sentences are not equivalent, but share some details; (1) The two sentences are not equivalent, but are on the same topic; (0) The two sentences are on different topics. To check agreement, the first two 20-pair samples were scored by a second French native speaker. Besides, in a few situations, the first judge was sometimes unsure whether to give a score or the next higher score. In these situations, he entered two alternate scores: this created a second series of judgments which differed only in a few places. Altogether, five batches were examined: three for VIC and two for RALI, and for each batch, we had two series of judgments.

For VIC, we sampled 60 sentence pairs from the 978 false positives of the most precise run, Run 2. Out of these sentence pairs, 3–5 were considered as perfect translations (grade 5) and an additional 8–13 were judged as near-perfect translations (grade 4).

From this we computed four increasingly lenient evaluations based upon the minimum and maximum numbers of perfect translations (*5 min*, *5 max*) and upon the minimum and maximum numbers of perfect or near-perfect translations (*4–5 min*, *4–5 max*). We converted these counts into percentages of the examined false positives that were judged as true translations (*T%FP*). We then

extrapolated these percentages to the whole set of false positives to obtain the number of human-judged true positives that should be added to the automatically evaluated true positives (*+TP*). We used this additional number to recompute the true positives and the associated precision (*P'*). Recall cannot be recomputed this way, because to estimate the recall for both automatic and 'natural' translation pairs, we would need to draw a sample from the full test corpus, and given the low prevalence of 'natural' translation pairs, this sample should be quite large. Table 6 shows the corresponding evolution of precision. For information we also recomputed the F1-score (*F1'*, still without changing the recall). We observe that precision is reevaluated with an increase of up to 4pt, whereas F1-score gains up to 1.5pt. This difference cannot be ignored for a precise evaluation, but does not bring drastic changes to the overall conclusions of the shared task.

Evaluation	T%FP	+TP	P' (%)	F1' (%)
base (auto)	0.0	0	87.1	79.4
5 (min)	5.0	49	87.7	79.6
5 (max)	8.3	82	88.2	79.8
4–5 (min)	18.3	179	89.4	80.3
4–5 (max)	30.0	293	91.0	80.9

Table 6: Re-evaluation of precision for VIC's Run 2. 'T%FP' is the percentage of human-assessed good translations in the false positives.

For RALI, we sampled 40 sentence pairs from the 41,865 false positives of the most precise run, Run 2. Out of these sentence pairs, none was considered as perfect translations nor near-perfect translations (most were related though). This is consistent with the fact that RALI2's precision was seven times lower than that of VIC2: a much larger sample might be needed to evidence 'natural' translation pairs in RALI2's output.

This limited experiment suggests that 'natural' translation pairs are much less frequent in the French-English test set than our artificially inserted translation pairs (or that the VIC2 system is much better at spotting the inserted translation pairs than 'natural' translation pairs): Table 6 shows that out of 7569 sentence pairs proposed by VIC2, 87% were inserted translation pairs and between 0.6% and 4% were 'natural' translation pairs. This would extrapolate to a rate of less than 5% of 'natural' translation pairs among the total translation pairs in the corpus.

An important limitation of this experiment is that it examined only a limited sample of sentence pairs, which entails large confidence intervals around the reported values. To compute these confidence intervals, we would need to know more or to make hypotheses about the distribution of 'natural' translation pairs not only in the system-returned sets of sentence pairs, but also outside these sets, which would require more time.

6 Conclusion

We presented the design and results of the second BUCC 2017 Shared Task, which consisted in spotting parallel sentences in comparable corpora. Some participants proposed creative methods, and the best results are quite high, with precisions, recalls and F1-scores between 80% and 88% depending on the language pair. The participants' papers contain directions for further improvement of their methods and results.

To alleviate the need for costly human evaluation, we designed a dataset in which known parallel sentence pairs have been inserted into monolingual corpora. Two risks were associated with this strategy. First, some participants might have tried to 'game' the task by attempting to discover the inserted sentences, for instance using plagiarism detection methods; we are glad that no participant seems to have done so. Second, whereas we could control the inserted translation pairs and try to reduce the likelihood of occurrence of 'natural' translation pairs, we could not fully prevent some from occurring; human examination of sample results from the best runs suggests that 'natural' translation pairs add only a few percents to the inserted translation pairs, confirming the overall relevance of the BUCC 2017 Shared Task dataset and evaluation.

The BUCC 2017 Shared Task dataset and evaluation program can be downloaded from the shared task's Web page.[5]

Acknowledgments

We thank the participants for the time they invested in this task and Léonard Zweigenbaum for his help in assessing the French-English translations. This work was partially funded by the European Union's Horizon 2020 Marie Sklodowska

[5] https://comparable.limsi.fr/bucc2017/bucc2017-task.html.

Curie Innovative Training Networks—European Joint doctorate (ITN-EJD) Under grant agreement No:676207 (MiRoR). Part of this work was supported by a Marie Curie Career Integration Grant within the 7th European Community Framework Programme.

References

Sadaf Abdul-Rauf and Holger Schwenk. 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*. Association for Computational Linguistics, Stroudsburg, PA, USA, BUCC '09, pages 46–54. http://dl.acm.org/citation.cfm?id=1690339.1690351.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 497–511. http://www.aclweb.org/anthology/S16-1081.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Vancouver, Canada, pages 46–50. http://www.aclweb.org/anthology/W/W17/W17-1009.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics, Berlin, Germany, pages 554–563. http://www.aclweb.org/anthology/W16-2347.

Thierry Etchegoyhen and Andoni Azpeitia. 2016. Set-theoretic alignment for comparable corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2009–2018. http://www.aclweb.org/anthology/P16-1189.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*. Lille, France, volume 37 of *JMLR Workshop and Conference Proceedings*.

Francis Grégoire and Philippe Langlais. 2017. BUCC 2017 Shared Task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Vancouver, Canada, pages 51–55. http://www.aclweb.org/anthology/W/W17/W17-1010.

Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2017. BUCC2017: A hybrid approach for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Vancouver, Canada, pages 61–64. http://www.aclweb.org/anthology/W/W17/W17-1012.

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*. Association for Computational Linguistics, Boston, Massachusetts, USA, pages 265–272.

Martin Potthast, Tim Gollub, Matthias Hagen, Johannes Kiesel, Maximilian Michel, Arnd Oberländer, Martin Tippmann, Alberto Barrón-Cedeño, Parth Gupta, Paolo Rosso, and Benno Stein. 2012. Overview of the 4th international competition on plagiarism detection. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. CEUR-WS.org, volume 1178 of *CEUR Workshop Proceedings*. http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-PotthastEt2012.pdf.

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

Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, pages 72–79. https://doi.org/10.3115/1075096.1075106.

Zheng Zhang and Pierre Zweigenbaum. 2017. zNLP: Identifying parallel sentences in Chinese-English comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*. Association for Computational Linguistics, Vancouver, Canada, pages 56–60. http://www.aclweb.org/anthology/W/W17/W17-1011.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2016. Towards preparation of the second BUCC shared task: Detecting parallel sentences in comparable corpora. In *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 38–43. https://comparable.limsi.fr/bucc2016/pdf/BUCC08.pdf.