Overview of the Third 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 2018 shared task on parallel sentence extraction from comparable corpora. This task used the same data as the BUCC 2017 shared task. 17 runs were submitted by 3 teams, covering all four proposed language pairs: German-English (3 runs), French-English (6 runs), Russian-English (3 runs), and Chinese-English (5 runs). The best F-scores as measured against the gold standard were 0.86 (German-English), 0.81 (French-English and Russian-English), and 0.77 (Chinese-English). All top scores improved over those of 2017.

Keywords: Comparable corpora, parallel sentences, parallel sentence extraction, cross-language similarity, annotated corpus

1. Introduction

Comparable corpora are gaining momentum as a supplement to parallel corpora for multilingual natural language processing (Sharoff et al., 2013; Rapp et al., 2016). After the extraction of word translations (Rapp, 1995; Fung, 1995), the detection of parallel sentences (Utiyama and Isahara, 2003; Munteanu et al., 2004; Abdul Rauf and Schwenk, 2009a) and parallel segments (Munteanu and Marcu, 2006; Hewavitharana and Vogel, 2011) in comparable corpora was addressed and found to improve statistical machine translation (Munteanu and Marcu, 2005; Abdul-Rauf and Schwenk, 2009b).

This strong interest in comparable corpora created a need for shared tasks that provide common task definitions, datasets and evaluation methods to assess the state of the art. Such shared tasks were created in the context of the BUCC workshop series on Building and Using Comparable Corpora and in other venues: the first one was run at BUCC 2015 and addressed the detection of comparable documents in two languages (Sharoff et al., 2015). It was followed on the same topic by the bilingual document alignment task of WMT 2016 (Buck and Koehn, 2016). A task on parallel sentence extraction from comparable corpora was prepared in 2016 (Zweigenbaum et al., 2016) and organized at BUCC 2017 (Zweigenbaum et al., 2017). It bears relations with but differs in several respects from the cross-language plagiarism detection tasks of PAN (Potthast et al., 2012) and the cross-language semantic text similarity task of SemEval (Agirre et al., 2016).

To let more participants take part in this task, we decided to run it for a second year in 2018 as the Third BUCC Shared Task.[1]. In this paper we describe the task and its datasets (Section 2.), the participants' systems (Section 3.), the results they obtained (Section 4.), and conclude (Section 5.).

[1] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

2. Task and Datasets

As in the Second BUCC Shared Task, the Third BUCC Shared Task aims to examine the ability of algorithms to detect parallel sentence pairs in a pair of monolingual corpora. Its design principles are the following.

Observing that past work took advantage of much existing meta-information, such as links between two matching Wikipedia articles in two languages or article dates in synchronous comparable news corpora (Munteanu and Marcu, 2005), we decided to create a dataset in which algorithms should focus on sentence contents instead of trying to rely on external, contextual clues. This should remove a large part of the heuristic aspects of these algorithms that are not directly linked to detecting cross-language sentence parallelism. Therefore this BUCC dataset has no meta-information attached to documents or sentences. To prevent participants from obtaining such meta-information indirectly, the instructions asked them not to use the original datasets from which the BUCC dataset was built.

The main difficulty in preparing a dataset to evaluate parallel sentence extraction from a pair of comparable corpora is the preparation of gold standard annotations: these annotations must identify the true positive parallel sentence pairs among the much larger set of true negatives, i.e., non-parallel sentence pairs, among the cross-product of sentences of the two corpora. Because the cross-product grows with the product of the sizes of the two corpora, as soon as these sizes exceed a few hundred sentences, it becomes difficult, not to say impossible, to manually spot the few parallel sentence pairs that happen to occur in these comparable corpora.

We therefore designed a dataset in which (i) parallel sentence pairs have been artificially inserted, in a way to make their presence as inconspicuous as possible; and (ii) action has been taken to make naturally occurring parallel sentence pairs less likely to occur. More detail is provided in (Zweigenbaum et al., 2016; Zweigenbaum et al., 2017).

The dataset for the BUCC'18 shared task consits of two parts. The non-parallel part is made of Wikipedia sen-

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 (reproduced from (Zweigenbaum et al., 2017)): 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 (reference)	Language pairs (*-en)
H2@BUCC2018	Carnegie Mellon University in Qatar, Qatar & QCRI, Qatar (Bouamor and Sajjad, 2018)	fr (3)
NLP2CT	NLP2CT Lab, Dept. of Computer and Information Science, University of Macau (Leong et al., 2018)	zh (2)
VIC	Vicomtech-IK4, Donostia / San Sebastian, Gipuzkoa, Spain (Azpeitia et al., 2018)	de (3), fr (3), ru (3), zh (3)

Table 2: Shared task systems: system label, team affiliation, publication reference, number of runs for each language pair

tences (dumps as of 20161201[2]) in two chosen languages. The parallel part is made of News Commentary sentences (v11[3]). As mentioned above, the instructions required task participants not to use any of these two corpora in their methods and systems. Datasets were prepared for four language pairs, each of which included English and another language among German (de), French (fr), Russian (ru), and Chinese (zh). Each dataset contained sample, training, and test splits (see Table 1).

Given a dataset containing two monolingual corpora en and fr, systems were expected to produce a set of sentence pairs (s^i_{en}, s^i_{fr}). Evaluation was performed by comparing system pairs to the set of gold standard pairs, and computing precision, recall, and F1-score in the usual way.

Note that the gold standard was defined by artificially inserted sentences. There is however a non-zero chance that some other pairs of sentences naturally happen to be translations too. If a system finds such correct sentence pairs that are not part of the gold standard annotations, these pairs are counted as false positives. As a result, the precision of system runs can be underestimated. By reviewing a small sample of false positive sentence pairs in the most precise en-fr run of one of the Second BUCC Shared Task participants (Zweigenbaum et al., 2017), we computed a very rough estimate of the number of such sentence pairs. We considered as correct translations sentence pairs such that (i) "the two sentences are completely equivalent, as they mean the same thing," possibly also considering cases in which (ii) "the two sentences are mostly equivalent, but some unimportant details differ." These correspond to the top two grades (5 and 4) in the guidelines of cross-language sentence similarity in SemEval 2016 (Agirre et al., 2016). Lower grades, e.g. (3) in which "the two sentences are roughly equivalent, but some important information differs or is missing" were not considered correct translations. Table 3 lists examples of sentence pairs considered false positives according to the gold standard, together with the human judgments (s) they received. Two sentence pairs in Table 3 received different scores from the two judges.

fr	en	s
Le renforcement de la gendarmerie locale par des troupes européennes est vite envisagé.	The reinforcement of the local gendarmerie with European troops was quickly planned.	5
Avant la *Première Guerre mondiale*, l'Allemagne importait annuellement pour 1,5 milliard de Reichsmarks de matières premières en provenance de Russie.	Germany imported 1.5 billion Rechsmarks of raw materials *and other goods* annually from Russia before the war.	4, 5
Le Mozambique est l'un des pays les plus pauvres du monde.	Mozambique is one of the poorest *and most underdeveloped* countries in the world.	4
Le jeu comporte aussi *plusieurs* modes de jeu, qui peuvent être joué en solo ou en multijoueur local:	Competitive multiplayer modes have also *been added*, and can be played locally or over a network.	3, 4
Dans le deuxième, le type cystovarien, les ovocytes sont transmis à l'extérieur, par le biais de l'oviducte.	In the third type, the oocytes are conveyed to the exterior through the oviduct.	3

Table 3: Example sentence pairs found in false positive system output, with associated human cross-language similarity scores s. Italics emphasize extra material

We found that the resulting understimate of precision for that participant was between 0.6 and 4 points depending on whether only grade 5 pairs were considered correct, whether grade 4 pairs were also deemed acceptable, and on

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

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

how discordances across annotators were reconciled. Participants with less precise results were less subject to this phenomenon, therefore this did not change rankings.

3. Participants and systems

16 teams downloaded datasets, among which three teams submitted runs. Table 2 gives more detail about teams and runs.

Systems addressed the bilingual dimension of the task with machine translation systems (H2@BUCC2018, nlp2ct2), or used parallel corpora to obtain word translations (VIC) or to train bilingual word embeddings (H2@BUCC2018) or an autoencoder (nlp2ct2).

Cross-language sentence similarity was handled by the Jaccard coefficient (VIC) or the BLEU score (H2@BUCC2018), possibly with weighting (a function of frequency: VIC) and with a trained classifier (H2@BUCC2018, nlp2ct2).

One team used an Information Retrieval engine for faster search of similar sentences (VIC), where as the others took advantage of the fast computation of the Cosine of word embeddings (H2@BUCC2018) or of the orthogonal denoising encoder output (nlp2ct2).

4. Results and discussion

We present evaluation results for the runs submitted for each language. In each table we show the precision, recall and F1-score of each run in percentages. In addition, we show the best run of 2017 when available for that language pair. Because the evaluation performed through this synthetic dataset, with artificially inserted translation pairs, only approximates what a human evaluation of system results would return, it would not be relevant to compute scores with many digits: therefore we round the computed figures to the nearest integer.

Table 4 shows results for the three runs submitted on the German-English (de-en) language pair (one team). As in 2017, this language pair obtains the best results. Table 5 presents the six runs submitted on the French-English (fr-en) language pair by two teams. Table 6 presents the three runs submitted on the Russian-English (ru-en) language pair by one team. This language pair did not receive any submissions in 2017. Table 7 presents the five runs submitted on the Chinese-English (zh-en) language pair by two teams. They all improve upon the previous year's zh-en results.

5. Conclusion

The third BUCC 2018 Shared Task addressed spotting parallel sentences in comparable corpora. The best results of the participants are high, with precisions of 89–91%, recalls of 75–84%, and F1-scores of 77–86%. The Russian-English language pair was attempted for the first time, and the Chinese-English language pair was again the most challenging. F1-scores improved over 2017 for all language pairs. The BUCC 2018 Shared Task dataset and evaluation program can be downloaded from the shared task's Web page.[4]

run_name	sys_n	P	R	F1
VIC1.de-en	9271	87	**84**	**86**
VIC3.de-en	8265	**91**	79	85
VIC2.de-en	8769	88	81	84
VIC1.de-en in 2017	8640	88	80	**84**

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

run_name	sys_n	P	R	F1
VIC1.fr-en	8136	86	77	**81**
VIC2.fr-en	7173	**91**	72	80
VIC3.fr-en	8887	80	**79**	80
H2@BUCC18_1_fr-en	7947	82	72	76
H2@BUCC18_2_fr-en	9607	71	75	73
H2@BUCC18_3_fr-en	8300	70	64	67
VIC1.fr-en in 2017	8831	80	79	**79**

Table 5: Evaluation (%) of fr-en runs (n_gold=9,043).

run_name	sys_n	P	R	F1
VIC1.ru-en	11010	86	77	**81**
VIC2.ru-en	10127	**90**	71	79
VIC3.ru-en	11370	79	**79**	79

Table 6: Evaluation (%) of ru-en runs (n_gold=14,330)

Acknowledgments

We thank the participants for their interest in this task. This work was partially funded by the European Union's Horizon 2020 Marie Sklodowska 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.

6. References

Abdul Rauf, S. and Schwenk, H. (2009a). 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*, pages 46–54, Singapore, August. Association for Computational Linguistics.

Abdul-Rauf, S. and Schwenk, H. (2009b). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece, March. Association for Computational Linguistics.

Agirre, E., Banea, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Mihalcea, R., Rigau, G., and Wiebe, J. (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)*, pages 497–511, San Diego,

[4] https://comparable.limsi.fr/bucc2018/bucc2018-task.html.

run_name	sys_n	P	R	F1
VIC1.zh-en	1680	80	71	75
VIC2.zh-en	1373	**89**	64	74
VIC3.zh-en	1763	80	**75**	**77**
nlp2ct1.zh-en	1169	73	45	55
nlp2ct2.zh-en	1209	72	46	56
zNLP1 in 2017	1985	42	**44**	**43**

Table 7: Evaluation (%) of zh-en runs (n_gold=1,896)

California, June. Association for Computational Linguistics.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2018). Extracting parallel sentences from comparable corpora with STACC variants. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.

Bouamor, H. and Sajjad, H. (2018). H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.

Buck, C. and Koehn, P. (2016). Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany, August. Association for Computational Linguistics.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183.

Hewavitharana, S. and Vogel, S. (2011). Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 61–68, Portland, Oregon, June. Association for Computational Linguistics.

Leong, C., Wong, D. F., and Chao, L. S. (2018). UMpAligner: Neural network-based parallel sentence identification model. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, Miyazaki, Japan, May. ELRA.

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

Munteanu, D. S. and Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia, July. Association for Computational Linguistics.

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., 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.

Rapp, R., Sharoff, S., and Zweigenbaum, P. (2016). Recent advances in machine translation using comparable corpora. *Natural Language Engineering*, 22(4):501–516, July.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, student session*, volume 1, pages 320–322, Boston, Mass.

Sharoff, S., Rapp, R., and Zweigenbaum, P. (2013). Overviewing important aspects of the last twenty years of research in comparable corpora. In Serge Sharoff, et al., editors, *Building and Using Comparable Corpora*, pages 1–20. Springer, Berlin Heidelberg, December.

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.

Zweigenbaum, P., Sharoff, S., and Rapp, R. (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*, pages 38–43, PortoroÅŸ, Slovenia, May. European Language Resources Association (ELRA).

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada, August. Association for Computational Linguistics.