Extracting Parallel Sentences from Comparable Corpora with STACC Variants

Andoni Azpeitia, Thierry Etchegoyhen and Eva Martínez Garcia
Vicomtech, Donostia/San Sebastián, Spain
{aazpeitia, tetchegoyhen, emartinez}@vicomtech.org

Abstract

This article describes our submissions to the BUCC 2018 shared task on parallel sentence extraction from comparable corpora. Our approach is based on variants of the STACC method, which computes similarity on expanded lexical sets via Jaccard similarity. We apply the weighted variant of the method to all four language pairs of the task, demonstrating the efficiency and portability of the approach. Additionally, we introduce a variant which further penalizes mismatches in terms of named entities, improving over the already strong weighted variant baseline in most cases. Our approach reached the highest results in all scenarios, with scores over 80% in terms of f1-measure and 90% in precision.

Keywords: BUCC 2018, Shared Task, Sentence Alignment, Comparable Corpora

1. Introduction

The exploitation of comparable corpora is an important research area (Munteanu and Marcu, 2005; Sharoff et al., 2016), as it contributes to the creation of the parallel corpora that are needed for multilingual natural language processing tasks such as data-driven machine translation (Brown et al., 1990; Bahdanau et al., 2015) or automated bilingual dictionary creation (Rapp, 1995).

Extracting parallel sentences from comparable corpora is a challenging task, which has given rise to the development of a wide range of approaches over the years. Thus, interesting results have been notably obtained with methods based on suffix trees (Munteanu and Marcu, 2002), maximum likelihood (Zhao and Vogel, 2002), binary classification (Munteanu and Marcu, 2005), cosine similarity (Fung and Cheung, 2004), reference metrics over statistical machine translations (Abdul-Rauf and Schwenk, 2009; Sarikaya et al., 2009), feature-based approaches (Ştefănescu et al., 2012; Smith et al., 2010) or deep learning with bidirectional recurrent neural networks (Grégoire and Langlais, 2017), among others.

For our participation in the BUCC 2018 shared task on extracting parallel sentences from comparable corpora, we followed the STACC approach of (Etchegoyhen et al., 2016; Etchegoyhen and Azpeitia, 2016), which is based on Jaccard similarity (Jaccard, 1901) over lexical sets, with additional set expansion operations to address named entities and morphological variation.

We selected as our baseline the weighted variant of the approach (Azpeitia et al., 2017), which proved highly successful on the BUCC 2017 shared task (Zweigenbaum et al., 2017), and applied the approach to all four language pairs in the 2018 task. Additionally, we designed a variant of this approach which further penalizes mismatches in terms of named entities, showing that it improves over the strong weighted STACC baseline in most cases.

The results obtained in this shared task confirm the efficiency and portability of our approach, and additionally demonstrate the specific importance of named entities for parallel sentence extraction from comparable corpora.

2. STACC

The STACC approach has been described and explored in detail in (Etchegoyhen and Azpeitia, 2016), and we briefly summarise below how similarity is computed with their method.

Let s_i and s_j be two tokenised and truecased sentences in languages l_1 and l_2, respectively, S_i the set of tokens in s_i, S_j the set of tokens in s_j, T_{ij} the set of lexical translations into l_2 for all tokens in S_i, and T_{ji} the set of lexical translations into l_1 for all tokens in S_j.

Lexical translations are initially computed from sentences s_i and s_j by retaining the k-best translations for each word, if any, as determined by the ranking obtained from the lexical translation probabilities computed with IBM word alignment models (Brown et al., 1990). The sets T_{ij} and T_{ji} that comprise these k-best lexical translations are then expanded by means of two operations:

1. For each element in the set difference $T'_{ij} = T_{ij} - S_j$ (respectively $T'_{ji} = T_{ji} - S_i$), and each element in S_j (respectively S_i), if both elements share a common prefix with minimal length of more than n characters, the prefix is added to both sets. This longest common prefix matching strategy is meant to capture morphological variation via minimal computation.

2. Numbers and capitalised truecased tokens not found in the translation tables are added to the expanded translation sets. This operation addresses named entities, which are strong indicators of potential alignment given their low relative frequency and are likely to be missing from translation tables trained on different domains.

With source and target sets as defined here, the STACC similarity score is then computed as in Equation 1:

$$stacc(s_i, s_j) = \frac{\frac{|T_{ij} \cap S_j|}{|T_{ij} \cup S_j|} + \frac{|T_{ji} \cap S_i|}{|T_{ji} \cup S_i|}}{2} \quad (1)$$

Similarity for the core metric is thus defined as the average of the Jaccard similarity coefficients obtained between sentence token sets and expanded lexical translations in both directions.

2.1. STACC$_w$

In (Azpeitia et al., 2017), the STACC$_w$ variant of the core method is described, where set membership values of 1 in the original approach are replaced with lexical weights. The weights are computed according to Equation 2, where $f(w_i)$ is the relative frequency of word w_i and α is a parameter controlling the smoothness of the curve.

$$W(w_i) = \frac{1}{e^{\sqrt{\alpha \cdot f(w_i)}}} \tag{2}$$

Weighting can be computed on each monolingual corpus to be aligned, as will be the case for all the results reported in this paper, or on separate monolingual corpora. STACC$_w$ similarity is computed according to the weighted Jaccard similarity formula described in Equation 3, for a given lexical translation set T and token set S:

$$WJ(T,S) = \frac{\sum\limits_{w_m \in \{T \cap S\}} W(w_m)}{\sum\limits_{w_n \in \{T \cup S\}} W(w_n)} \tag{3}$$

The complete weighted similarity score is thus computed according to Equation 4.

$$stacc_w(s_i, s_j) = \frac{WJ(T_{ij}, S_j) + WJ(T_{ji}, S_i)}{2} \tag{4}$$

This variant was rather successful on the BUCC 2017 shared task, as it significantly improved over the baseline version of STACC, which would have already obtained the best results on all metrics in the two language pairs alignment scenarios in which the system participated.

2.2. STACC$_{wp}$

For this version of the BUCC shared task, we introduced a new variant, based on STACC$_w$ and on a penalty oriented towards named entity mismatches.

Both STACC and STACC$_w$ include a treatment of named entities, defined in terms of surface forms, by including in the expanded translation sets both capitalised words and numbers. Intuitively though, named entities might be thought of as playing an even stronger role than simply participating in determining similarity: when glancing over sets of comparable sentences, checking mismatches in terms of named entities between a given pair of sentences seems an efficient method to at least quickly discard improbable alignments.

We tested this hypothesis by first defining a penalty as in Equation 5, where N_i and N_j denote the sets of surface-form entities in the source and target sentence, respectively.

$$nep(s_i, s_j) = \frac{|(N_i - N_j) \cup (N_j - N_i)|}{|S_i \cup S_j|} \tag{5}$$

The penalty is thus defined in terms of set differences, taking as numerator the union of entities that are present in one sentence but not in the other. By defining the denominator as the union of all tokens in the source and target sentences, the measure is bound between 0 and 1, and a higher penalty will be assigned to sentence pairs with larger numbers of mismatching entities.

For this STACC$_{wp}$ variant, the penalty is included in the computation of the final score according to Equation 6.

$$stacc_{wp}(s_i, s_j) = stacc_w(s_i, s_j) - nep(s_i, s_j) \tag{6}$$

Thus, this variant preserves the successful core weighted metric for all cases where either no entities are present in the source and target sentences, or when the same entities are present in both sentences. The penalty complements the core metric by gradually reducing the overall score as entity mismatches increase between the source and target sentences.

3. BUCC 2018 Shared Task

The BUCC 2018 shared task on parallel sentence extraction from comparable corpora[1] consists in identifying translation pairs within two sentence-split monolingual corpora. It involves four language pairs and we applied the variants of our approach in all four alignment scenarios. The organisers provided three datasets for each language pair, whose statistics are described in Table 1; gold reference pairs were provided for the training and sample sets.

3.1. Experimental Settings

The volumes of data selected for the task makes it unrealistic to compute the alignments over the Cartesian products of source and target sentences. Thus, we use the STACC system in cross-language information retrieval (CLIR) mode, where target sentences are first indexed using the Apache Lucene toolkit [2] and retrieved by building a query over the expanded sets created from each source sentence.

This strategy drastically reduces the computational load, at the cost of missing some correct alignment pairs. Similarity is computed for each source sentence against all retrieved candidates and a final optimisation is applied to enforce 1-1 alignments, a process which has been shown to improve the quality of alignments (Etchegoyhen and Azpeitia, 2016).

For each language pair, weighting was computed on each monolingual corpus composing the pair to be aligned. Translation tables were generated with the GIZA++ toolkit (Och and Ney, 2003) for all language pairs but Russian-English, for which word alignments were computed with FastAlign (Dyer et al., 2013).

To train the word alignment models, we followed the approach in (Azpeitia et al., 2017) and created generic corpora via bilingual perplexity-based sampling, with an arbitrary upper data selection bound to avoid over-representing individual corpora. Note that, due to time availability to prepare our submissions, this method was not applied to our two new language pairs, Russian-English and Chinese-English, for which we only used the MULTIUN corpus, in totality for the former, and a sample of approximately 2 million for the latter. Table 2 describes the number of sentence pairs selected for each language pair.[3]

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

[2] https://lucene.apache.org.

[3] All original corpora were downloaded from the OPUS repository (Tiedemann, 2012): http://opus.lingfil.uu.se/; the upper selection bound was set to 500,000 sentence pairs after considering the relative weights of the available corpora.

PAIR	LANG	MONOLINGUAL			GOLD		
		SAMPLE	TRAIN	TEST	SAMPLE	TRAIN	TEST
DE-EN	de	32,593	413,869	413,884	1,038	9,580	9,550
	en	40,354	399,337	396,534	1,038	9,580	9,550
FR-EN	fr	21,497	271,874	276,833	929	9,086	9,043
	en	38,069	369,810	373,459	929	9,086	9,043
RU-EN	ru	45,459	460,853	457,327	2,374	14,435	14,330
	en	72,766	558,401	566,356	2,374	14,435	14,330
ZH-EN	zh	8,624	94,637	91,824	257	1,899	1,896
	en	13,589	88,860	90,037	257	1,899	1,896

Table 1: Task data statistics (number of sentences)

PAIR	DATA	CORPUS					
		OPENSUBS	MULTIUN	EUROPARL	JRC	TED	GENERIC
DE-EN	Original	11,473,328	103,490	1,776,292	449,818	138,243	*13,941,171*
	Selected	500,000	103,490	500,000	449,818	139,243	*1,692,551*
FR-EN	Original	28,024,360	9,142,161	1,826,770	708,896	153,167	*39,855,354*
	Selected	500,000	500,000	500,000	316,327	153,167	*1,969,494*
RU-EN	Original	-	9,111,212	-	-	-	*9,111,212*
	Selected	-	9,111,212	-	-	-	*9,111,212*
ZH-EN	Original	-	7,747,328	-	-	-	*7,747,328*
	Selected	-	1,831,016	-	-	-	*1,831,016*

Table 2: Generic data (number of sentences)

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.15	99.04	95.09	**91.51**	**93.27**
SAMPLE	STACC$_{wp}$ (F)	250	0.15	99.04	97.36	89.01	93.00
SAMPLE	STACC$_{wp}$ (P)	250	0.16	99.04	**99.21**	85.54	91.87
TRAIN	STACC$_w$ (F)	250	0.17	98.50	87.00	79.96	83.33
TRAIN	STACC$_{wp}$ (F)	250	0.16	98.50	84.81	**83.74**	**84.27**
TRAIN	STACC$_{wp}$ (P)	250	0.17	98.50	**89.86**	78.28	83.67
TEST	STACC$_w$ (F)	250	0.17	98.65	88.06	80.86	84.31
TEST	STACC$_{wp}$ (F)	250	0.16	98.65	86.81	**84.27**	**85.52**
TEST	STACC$_{wp}$ (P)	250	0.17	98.65	**91.47**	79.16	84.87

Table 3: Results for DE-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.15	99.46	92.44	89.45	90.92
SAMPLE	STACC$_{wp}$ (F)	250	0.14	99.46	92.26	**91.07**	**91.66**
SAMPLE	STACC$_{wp}$ (P)	250	0.15	99.46	**95.33**	87.84	91.43
TRAIN	STACC$_w$ (F)	250	0.16	96.84	78.43	**79.23**	78.83
TRAIN	STACC$_{wp}$ (F)	250	0.16	96.84	83.93	77.58	**80.63**
TRAIN	STACC$_{wp}$ (P)	250	0.17	96.84	**87.81**	71.69	78.93
TEST	STACC$_w$ (F)	250	0.16	96.87	80.27	**78.89**	79.58
TEST	STACC$_{wp}$ (F)	250	0.16	96.87	86.01	77.39	**81.47**
TEST	STACC$_{wp}$ (P)	250	0.17	96.87	**90.62**	71.88	80.17

Table 4: Results for FR-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.12	100.00	91.27	**89.49**	**90.37**
SAMPLE	STACC$_{wp}$ (F)	250	0.12	100.00	95.79	70.82	81.43
SAMPLE	STACC$_{wp}$ (P)	250	0.13	100.00	**98.82**	65.37	78.69
TRAIN	STACC$_w$ (F)	250	0.14	97.05	78.27	**74.72**	**76.45**
TRAIN	STACC$_{wp}$ (F)	250	0.13	97.05	79.26	70.62	74.69
TRAIN	STACC$_{wp}$ (P)	250	0.14	97.05	**86.23**	64.61	73.87
TEST	STACC$_w$ (F)	250	0.14	97.15	80.37	**74.74**	**77.45**
TEST	STACC$_{wp}$ (F)	250	0.13	97.15	79.82	70.73	75.00
TEST	STACC$_{wp}$ (P)	250	0.14	97.15	**88.64**	64.19	74.46

Table 5: Results for ZH-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
SAMPLE	STACC$_w$ (F)	250	0.15	97.81	95.42	86.98	91.01
SAMPLE	STACC$_{wp}$ (F)	250	0.14	97.81	96.46	**88.37**	**92.24**
SAMPLE	STACC$_{wp}$ (P)	250	0.15	97.81	**97.94**	84.16	90.53
TRAIN	STACC$_w$ (F)	250	0.16	96.64	77.69	**79.77**	78.72
TRAIN	STACC$_{wp}$ (F)	250	0.16	96.64	84.87	77.26	**80.89**
TRAIN	STACC$_{wp}$ (P)	250	0.17	96.64	**88.05**	71.02	78.63
TEST	STACC$_w$ (F)	250	0.16	96.81	79.44	**79.34**	79.39
TEST	STACC$_{wp}$ (F)	250	0.16	96.81	86.31	76.83	**81.30**
TEST	STACC$_{wp}$ (P)	250	0.17	96.81	**89.91**	70.67	79.14

Table 6: Results for RU-EN

Regarding STACC hyper-parameters, k-best lexical translations were limited to a maximum of 4 and the minimal prefix length for longest common prefix matching was set to 4. Lucene indexing was based on words with length of 4 or more characters, and a maximum of 100 candidates were retrieved for each source sentence. For each language pair, English was arbitrarily set to be the target language. For the weighting function, α was set to 250 across the board, as it was established in (Azpeitia et al., 2017) to be an optimal setting overall.

We prepared three variants for the task and applied all three on all four language pairs. The first variant is STACC$_w$, which we take to be our baseline, with an alignment threshold set to maximise the f1-measure on the training set. The second variant is the STACC$_{wp}$ method described in Section 2.2., with an alignment threshold also set to maximise the f1-measure.[4] Finally, we submitted a third variant, based on STACC$_{wp}$ but with a higher alignment threshold meant to maximise precision, as in practical cases it may be optimal to create smaller but more accurate bitexts from comparable corpora.[5]

3.2. Results

Results on all datasets are shown in Tables 3, 4, 5 and 6, along with the hyper-parameters used for each dataset and the percentage of correct candidates retrieved via Lucene indexing and retrieval. Our system competed with other systems in FR-EN and ZH-EN, with our variants reaching the highest scores on all three metrics;[6] for DE-EN and RU-EN, there were no other competing systems.

Since not all gold parallel sentences are known for this task, the results shown here are minimum values, i.e. there may be actually correct alignments identified as false positives.[7] They are nonetheless satisfactory across the board, with f1 scores above 80% on the test sets for French-English, German-English and Russian-English, and precision above 90% for the same three pairs. Although slightly lower, Chinese-English results are close to the 80% mark for the f1 measure and at 89% in terms of precision, improving over the best results obtained for this language pair on the similar BUCC 2017 task by more than 30 f1-measure points and over 40 points in terms of precision.

Our submission this year confirmed the efficiency of the generic STACC approach on Russian and Chinese, two languages that exhibit marked differences with the other two language pairs. Thus, these results further validate the claim of portability for our approach.

As for the STACC$_{wp}$ variant we introduced this year, it provided significant improvements over the already robust STACC$_w$ method, with gains of up to two points in f1-measure. Only for Chinese-English were the results lower than with STACC$_w$, a not completely unexpected result given the peculiarities of Chinese in terms of named entities as well. The results obtained with this variant confirm the specific importance of named entities for the alignment of comparable sentences, and the need to give them special prominence when computing alignment scores.

Overall, we view the high scores obtained on all metrics in all language pairs as satisfactory, especially considering the large test sets used in the shared task.

4. Conclusion

We described our submission to the 2018 BUCC shared task on the extraction of parallel sentences from comparable corpora. Our contribution for this year was twofold. We first applied our STACC$_w$ approach, which is based on weighted set-theoretic operations on expanded lexical sets, to all four language pairs proposed for the task. Additionally, we introduce a variant which further penalizes mismatches in terms of named entities, improving over the already strong weighted variant baseline in most cases. This variant is seamlessly integrated into STACC via a set-based penalty computed over surface-defined named entities.

Our approach reached the highest results on all metrics and in all scenarios, with scores over 80% in terms of f1-measure and 90% in precision. The results from our participation in the BUCC 2018 shared task thus demonstrate the efficiency of the STACC approach in terms of quality of extracted alignments and portability across language pairs.

[4] Note that, for the German-English pair, the penalty was computed with named entity sets that only comprised numbers, as including capitalised words would have also captured common nouns that are not part of the translation tables because of lexical coverage gaps in the corpora.

[5] In the tables, we add an (F) next to each variant name if the alignment threshold was selected to optimise the f1-measure, and a (P) if set for precision.

[6] This claim is based on the results provided by the organisers as of this writing, which include the maximum scores obtained for the task in terms of the three metrics.

[7] See (Zweigenbaum et al., 2017) for an analysis of the improved results obtained via a sample-based complementary human evaluation.

5. Acknowledgements

This work was partially supported by the Spanish Ministry of Economy and Competitiveness and the Department of Economic Development and Competitiveness of the Basque Government via projects AdapTA (RTC-2015-3627-7) and TRADIN (IG-2015/0000347). We would like to thank MondragonLingua Translation & Communication as coordinator of these projects for their support.

6. References

Abdul-Rauf, S. and Schwenk, H. (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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Azpeitia, A., Etchegoyhen, T., and Martinez García, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *In Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Etchegoyhen, T. and Azpeitia, A. (2016). Set-Theoretic Alignment for Comparable Corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pages 2009–2018.

Etchegoyhen, T., Azpeitia, A., and Pérez, N. (2016). Exploiting a Large Strongly Comparable Corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Fung, P. and Cheung, P. (2004). Mining Very Non-Parallel Corpora: Parallel Sentence and Lexicon Extraction via Bootstrapping and E.M. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 57–63.

Grégoire, F. and Langlais, P. (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*, pages 46–50. Association for Computational Linguistics.

Jaccard, P. (1901). Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241 – 272.

Munteanu, D. S. and Marcu, D. (2002). Processing Comparable Corpora With Bilingual Suffix Trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 289–295. Association for Computational Linguistics.

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

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322. Association for Computational Linguistics.

Sarikaya, R., Maskey, S., Zhang, R., Jan, E.-E., Wang, D., Ramabhadran, B., and Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In *Proceedings of InterSpeech*, pages 432–435.

Sharoff, S., Rapp, R., Zweigenbaum, P., and Fung, P. (2016). *Building and Using Comparable Corpora*. Springer Publishing Company, Incorporated, 1st edition.

Smith, J. R., Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ştefănescu, D., Ion, R., and Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 137–144.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference*, pages 2214–2218.

Zhao, B. and Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, pages 745–748. IEEE.

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. Association for Computational Linguistics.