Weighted Set-Theoretic Alignment of Comparable Sentences

Andoni Azpeitia and **Thierry Etchegoyhen** and **Eva Martínez Garcia**
Vicomtech-IK4
Mikeletegi Pasalekua, 57
Donostia / San Sebastian, Gipuzkoa, Spain
{aazpeitia, tetchegoyhen, emartinez}@vicomtech.org

Abstract

This article presents the STACC$_w$ system for the BUCC 2017 shared task on parallel sentence extraction from comparable corpora. The original STACC approach, based on set-theoretic operations over bags of words, had been previously shown to be efficient and portable across domains and alignment scenarios. We describe an extension of this approach with a new weighting scheme and show that it provides significant improvements on the datasets provided for the shared task.

1 Introduction

Parallel corpora are an essential resource for the development of multilingual natural language processing applications, in particular statistical and neural machine translation (Brown et al., 1990; Bahdanau et al., 2014). Since the professional translations that are necessary to build quality bitexts are expensive and time-consuming, the exploitation of monolingual corpora that address similar topics, known as comparable corpora, has been extensively explored in the last two decades (Munteanu and Marcu, 2005; Sharoff et al., 2016).

A critical part of the process when building parallel resources from comparable data is the alignment of sentences in monolingual corpora. Over the years, several methods have been developed and evaluated for this task, including maximum likelihood (Zhao and Vogel, 2002), suffix trees (Munteanu and Marcu, 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), and feature-based approaches (Ştefănescu et al., 2012; Smith et al., 2010), among others.

For comparable sentence alignment, we followed the STACC approach in (Etchegoyhen et al., 2016; Etchegoyhen and Azpeitia, 2016), which is based on seed lexical translations, simple set expansion operations and the Jaccard similarity coefficient (Jaccard, 1901). This method has been shown to outperform state-of-the-art alternatives on a large range of alignment tasks and provides a simple yet effective procedure that can be applied across domains and corpora with minimal adaptation and deployment costs.

In this paper, we describe STACC$_w$, an extension of the approach with a word weighting scheme, and show that it provides significant improvements on the datasets provided for the BUCC 2017 shared task, while maintaining the portability of the original approach.

2 STACC

STACC is an approach to sentence similarity based on expanded lexical sets and Jaccard similarity, whose main goal is to provide a portable and efficient alignment mechanism for comparable sentences. The similarity score is computed as follows.

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 IBM models.[1] Lexical translations are selected according to the ranking provided by the precomputed lexical probabilities, without using the

[1] Translation tables are generated with the GIZA++ toolkit (Och and Ney, 2003).

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 41–45,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

actual probability values in the computation of similarity. The sets T_{ij} and T_{ji} that comprise the 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.

No additional operations are performed on the created sets, and in particular no filtering is applied, with punctuation and functional words kept alongside content words in the final sets. With source and target sets as defined here, the STACC similarity score is then computed as in Equation 1:

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

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

For scenarios where the alignment space is large, target sentences are first indexed using the Lucene search engine[2] and retrieved by building a query over the expanded translation sets created from each source sentence. This strategy drastically reduces the computational load, at the cost of missing some correct alignment pairs. In this mode, one of the two corpora is set as source and the other as target, retrieving n target alignment candidates for each source sentence. Similarity is computed over all candidates and a final optimisation process is applied that enforces 1-1 alignments, a process which has been shown to improve the quality of alignments (Etchegoyhen and Azpeitia, 2016).

3 Weighted STACC

Although STACC has been shown to outperform competing state-of-the-art approaches on a variety of domains and scenarios (Etchegoyhen and Azpeitia, 2016), it ignores lexical weights and thus assigns equal importance to open-class and function words. Although it makes intuitive sense to assign different weights according to the information provided by each word, adequate lexical weighting for a given task is not straightforward. Standard approaches such as TF-IDF often need to be complemented with stop word lists, which can be large and difficult to determine in agglutinative languages, for instance. Term-based approaches in general might assign weights that are too unbalanced for the task at hand, and termhood might be dependent on building accurate contrastive generic corpora (Gelbukh et al., 2010).

We follow the empirical approach in (Mikolov et al., 2013), where the imbalance between frequent and rare words is controlled by a subsampling formula with two variables: an empirically determined threshold and word frequency. Experiments with their exact weighting scheme did not however provide optimal results for our alignment goals. We opted instead to compute lexical weights 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.

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

Among the methods we tested empirically, this function has properties that fit rather well the original STACC approach. First, since it is bound between zero and one, it preserves the idea that set membership is a fruitful factor to compute similarity. Secondly, it assigns weights close to 1 for most open-class words while not completely discarding functional words,[3] a feature which has provided optimal results in our experiments.

Weighting is computed on each monolingual corpus to be aligned, thus removing any dependence on defining contrastive generic corpora. STACC$_w$ similarity is then computed according to the previously defined equation, except that set membership values of 1 in the original approach are replaced with lexical weights.

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

[3] The most frequent words typically receive a weight around 0.1 in the various distributions we tested.

PAIR	LANG	MONOLINGUAL			GOLD	
		TRAIN	SAMPLE	TEST	TRAIN	SAMPLE
DE-EN	de	413,869	32,593	413,884	9,573	1,037
	en	399,337	40,354	396,534	9,573	1,037
EN-FR	fr	271,874	21,497	276,833	9,080	929
	en	369,810	38,069	373,459	9,080	929

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*

Table 2: Generic data (number of sentences)

4 BUCC 2017 Shared Task

The BUCC 2017 shared task on parallel sentence extraction from comparable corpora[4] consists in identifying translation pairs within two sentence-split monolingual corpora. It involves four language pairs, from which we selected French-English and German-English for our participation.

The organisers provided three datasets for each language pair, whose statistics are described in Table 1 for the two language pairs we selected; gold reference pairs were provided for the training and sample sets.

Note that the statistics shown here differ slightly from those of the original data provided by the organisers, as we removed the bilingual duplicates that were found.[5]

4.1 Experimental Settings

Both STACC and STACC$_w$ require lexical translation tables to compute similarity, the only external source of information needed in this approach. In previous work (Etchegoyhen and Azpeitia, 2016), GIZA tables had been created from the JRC corpora only. In order to extend lexical coverage, we opted for a different approach and created generic translation tables from varied corpora.

In each corpus, parallel sentence pairs were first sorted by increasing perplexity scores according to language models trained on the monolingual side of each parallel corpus, where the score was taken to be the mean of source and target perplexities. A portion of each corpus was then selected to compose the final corpus, with an upper selection bound taken to be either the median average perplexity score or the top n pairs if selecting up to median perplexity would result in over representing the corpus. Table 2 describes the number of sentence pairs selected for each language pair, the lexical translation tables being extracted from the GENERIC datasets.[6]

Regarding 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 set to be the target language. We experimented with different values of α to control the smoothness of the weighting function and different values for the alignment threshold th used to discard low-confidence alignments.

Since up to three different runs could be submitted for the task, we prepared three variants of the system, where parameters α and th were set according to the best f-measure, precision and recall scores, respectively, obtained on the training set.[7]

Each of these variants was submitted to the task, in order to evaluate the behaviour of our system when targeting for precision, recall and f-measure. Although not submitted to the shared task, the original STACC method was also evaluated on the train and sample sets.

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

[5] There were 7 and 1 duplicates in the train and sample sets, respectively, for DE-EN, and 6 in the FR-EN train set.

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

[7] We identify these variants with F, P and R upperscripts in the tables.

DATASET	SYSTEM	α	th	LUCENE	P	R	F
TRAIN	STACC$_w{}^{F}$	250	0.17	98.50	86.99	79.96	**83.33**
TRAIN	STACC$_w{}^{P}$	250	0.18	98.50	**90.89**	73.41	81.23
TRAIN	STACC$_w{}^{R}$	250	0.16	98.50	80.21	**85.55**	82.79
TRAIN	STACC	–	0.23	98.50	79.26	69.16	73.87
SAMPLE	STACC$_w{}^{F}$	100	0.16	99.04	95.46	91.32	**93.35**
SAMPLE	STACC$_w{}^{P}$	100	0.17	99.04	**97.95**	87.75	92.57
SAMPLE	STACC$_w{}^{R}$	100	0.15	99.04	88.27	**93.64**	90.88
SAMPLE	STACC	–	0.22	99.04	91.84	80.33	85.70
TEST	STACC$_w{}^{F}$	250	0.17	98.63	88.15	79.75	**83.74**
TEST	STACC$_w{}^{P}$	250	0.18	98.63	**92.10**	73.16	81.55
TEST	STACC$_w{}^{R}$	250	0.16	98.63	81.93	**85.35**	83.60

Table 3: Results for DE-EN

DATASET	SYSTEM	α	th	LUCENE	P	R	F
TRAIN	STACC$_w{}^{F}$	250	0.16	96.84	78.43	79.23	**78.83**
TRAIN	STACC$_w{}^{P}$	250	0.17	96.84	**84.36**	73.40	78.50
TRAIN	STACC$_w{}^{R}$	250	0.15	96.84	68.51	**83.83**	75.40
TRAIN	STACC	–	0.23	96.84	72.69	63.12	67.57
SAMPLE	STACC$_w{}^{F}$	500	0.14	99.46	90.51	91.39	**90.95**
SAMPLE	STACC$_w{}^{P}$	500	0.15	99.46	**93.74**	86.98	90.23
SAMPLE	STACC$_w{}^{R}$	500	0.13	99.46	83.13	**93.33**	87.93
SAMPLE	STACC	–	0.22	99.46	89.36	75.03	81.57
TEST	STACC$_w{}^{F}$	250	0.16	96.81	80.41	78.52	**79.46**
TEST	STACC$_w{}^{P}$	250	0.17	96.81	**87.08**	72.89	79.35
TEST	STACC$_w{}^{R}$	250	0.15	96.81	69.82	**83.14**	75.90

Table 4: Results for FR-EN

4.2 Results

Results on all datasets are shown in Tables 3 and 4, along with the parameters used for each dataset and the percentage of correct candidates retrieved via Lucene indexing and search. On the test sets, our system competed with four other systems in FR-EN and our three submitted variants obtained the best results on all three metrics; for DE-EN, there were no other competing systems.

Given the nature of the evaluation, where not all gold parallel sentences are known, pairs identified as false positives may actually be correct alignments.[8] The results shown here are therefore minimum values and the already high scores achieved by our approach were thus quite satisfactory.

Overall, STACC$_w$ improves significantly over its non-weighted variant on the training and sample datasets, with improvements of around 10 points in f-measure on the training and sample sets. On the smaller sample sets, the accuracy of the alignments was naturally higher, reaching f-measure minimum scores above the 90% mark.

As expected, each variant of the system was better on the measure it was meant to optimise via adjustments of the alignment threshold.

An interesting additional result, not shown in the tables, is the weak impact of the hyperparameter α: between 100 and 500, the scores were marginally different; only values markedly outside this range gave worse results. These results were consistent for both training and sample sets, showing that the weighting function appears not to need corpus-specific adjustments for this parameter, a welcome result on portability grounds.

5 Conclusion

We described STACC$_w$, a weighted set-theoretic alignment method to extract parallel sentences from comparable corpora, which was the top ranked system in the BUCC 2017 shared task on the datasets where it competed with other systems and achieved high minimum value scores across the board. Our approach features generic lexical translation tables, Jaccard similarity over simple expanded translation sets and a generic word weighting scheme. This method improved significantly over the previous non-weighted approach on the provided training and sample datasets, while maintaining its main goals of portability, efficiency and ease of deployment.

[8] A quick manual evaluation of a sample of false positives confirmed that many were in fact correct alignments.

Acknowledgments

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

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473* .

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics* 16(2):79–85.

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, ACL 2016, August 7-12, 2016, Berlin, Germany*. volume 1: Long Papers, pages 2009–2018.

Thierry Etchegoyhen, Andoni Azpeitia, and Naiara Pérez. 2016. Exploiting a Large Strongly Comparable Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Pascale Fung and Percy Cheung. 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.

Alexander Gelbukh, Grigori Sidorov, Eduardo Lavin-Villa, and Liliana Chanona-Hernandez. 2010. Automatic term extraction using log-likelihood based comparison with general reference corpus. In *International Conference on Application of Natural Language to Information Systems*. Springer, pages 248–255.

Paul Jaccard. 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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

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

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics* 29(1):19–51.

Ruhi Sarikaya, Sameer Maskey, R Zhang, Ea-Ee Jan, D Wang, Bhuvana Ramabhadran, and Salim Roukos. 2009. Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In *Proceedings of InterSpeech*. pages 432–435.

Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, and Pascale Fung. 2016. *Building and Using Comparable Corpora*. Springer Publishing Company, Incorporated, 1st edition.

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

Dan Stefănescu, Radu Ion, and Sabine Hunsicker. 2012. Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation*. pages 137–144.

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

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