Set-Theoretic Alignment for Comparable Corpora

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

Abstract

We describe and evaluate a simple method to extract parallel sentences from comparable corpora. The approach, termed STACC, is based on expanded lexical sets and the Jaccard similarity coefficient. We evaluate our system against state-of-the-art methods on a large range of datasets in different domains, for ten language pairs, showing that it either matches or outperforms current methods across the board and gives significantly better results on the noisiest datasets. STACC is a portable method, requiring no particular adaptation for new domains or language pairs, thus enabling the efficient mining of parallel sentences in comparable corpora.

1 Introduction

With the rise of data-driven machine translation, be it statistical (Brown et al., 1990), example-based (Nagao, 1984), or rooted in neural networks (Bahdanau et al., 2014), the need for large parallel corpora has increased accordingly. Although quality bitexts have been made available over the years (Tiedemann, 2012), creating parallel corpora is a resource-consuming effort involving professional human translation of large volumes of texts in multiple languages. As a consequence, there is still a lack of parallel data to properly model translation across languages and domains.

To overcome this limitation, special emphasis has been placed in the last two decades on the exploitation of comparable corpora, with the development of a range of methods to mine parallel sentences from texts addressing similar topics in different languages. The work we present follows this line of research, describing and evaluating a simple method that allows parallel sentences to be efficiently mined in different languages and domains with minimal adaptation effort.

The method we describe, termed STACC, is based on expanded lexical sets and the Jaccard similarity coefficient (Jaccard, 1901), which is computed as the ratio of set intersection over union. We evaluate this simple approach against state-of-the-art methods for comparable sentence alignment on a variety of datasets for ten different language pairs, showing that STACC either matches or outperforms competing approaches.

The paper is organised as follows: Section 2 describes related work on parallel sentence mining in comparable corpora; Section 3 presents the STACC method; Section 4 describes the experiments in comparable sentence alignment, including the description of test corpora and systems, and an analysis of the results; Section 5 presents results obtained with an optimised version of the alignment process, beyond system comparison; finally, Section 6 draws conclusions from the work described in the paper.

2 Related work

A large variety of techniques have been proposed to mine parallel sentences in comparable corpora. One of the first approaches was proposed by (Zhao and Vogel, 2002), who combined sentence length and bilingual lexicon models under a maximum likelihood criterion. (Munteanu and Marcu, 2002) explored the use of suffix trees, later opting for maximum entropy-based binary classification using a modified version of IBM Model 1 word translation probabilities (Brown et al., 1993)

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2009–2018,
Berlin, Germany, August 7-12, 2016.

and both general and alignment-specific features (Munteanu and Marcu, 2005). (Fung and Cheung, 2004) describe the first approach to tackle parallel sentence mining in very non-parallel corpora, using cosine similarity as their sentence selection criterion.

Several approaches have employed full statistical machine translation models instead of relying only on lexical tables. (Abdul-Rauf and Schwenk, 2009), for instance, apply the TER metric (Snover et al., 2006) on fully machine translated output to identify parallel sentences; (Sarikaya et al., 2009) use a similar approach but with BLEU (Papineni et al., 2002) as their similarity metric. One of the noted advantages of including full machine translation is the ability to better model the complex factors found in translation, e.g. fertility and contextual information, as compared to lexicon-based approaches. The latter enable, in principle, the capture of a larger set of lexical translation variants, and do not require the training of complete translation models.

Sophisticated feature-based approaches have been developed in recent years in order to provide a method that may apply to larger sets of language pairs and domains. (Stefănescu et al., 2012) report improvements over previous methods with a feature-based sentence similarity measure, an approach which is described in more detail in Section 4.2.1. Another feature-rich approach is described in (Smith et al., 2010), showing improvements over standard and improved binary classifiers; we describe their model in more details in Section 4.2.2.

Jaccard similarity, a core component of the approach we describe, has been standardly used as a text similarity measure in information retrieval and text summarisation tasks, or to compute semantic similarity (Pilehvar et al., 2013). For comparable corpora, it has been notably employed by (Paramita et al., 2013), who estimate document comparability by computing the coefficient on a subset of translated source sentences, discarding those containing large amounts of named entities or numbers, and taking the average of these sentence-level scores. The method we present in the next section builds on a related similarity measure as a direct indicator of comparable sentence similarity.

3 STACC

STACC is an approach to sentence similarity based on expanded lexical sets, whose main goal is to provide a simple yet effective procedure that can be applied across domains and corpora with minimal adaptation and deployment costs.

We start with the minimal set of bilingual information that can be automatically extracted from a seed parallel corpus, using lexical translations determined and ranked according to IBM models; word translations are computed in both directions using the GIZA++ toolkit (Och and Ney, 2003).

STACC relies on the Jaccard index, which defines set similarity as the ratio of set intersection over union. We base our comparable sentence similarity measure strictly on this index, applying it to expanded lexical sets as described below.

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 expanded translations into l_2 for all tokens in S_i, and T_{ji} the set of expanded translations into l_1 for all tokens in S_j. The STACC similarity score is then computed as in Equation 1:

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

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

The translation sets T_{ij} and T_{ji} are initially computed from sentences s_i and s_j by retaining the k-best lexical translations found in GIZA tables, if any. Lexical translations are selected according to the ranking provided by the precomputed lexical probabilities but the specific probability values are not used any further to compute similarity:[1] all potential translations are members of the translation set as tokens. Discarding this source of potentially exploitable information is mostly motivated by the relative reliability of lexical translation probabilities across domains. Lexical translations are usually extracted from a different domain than that of the comparable corpora at hand, typically using professionally created institutional corpora such as Europarl (Koehn, 2005), and lexical distributions across

[1] This differs from (Skadiņa et al., 2012), who include a lexical translation feature where actual probabilities are used to compute the final score.

domains can be expected to be quite different. This casts doubt on the usefulness of using precomputed translation probabilities and simple set membership was favoured in our approach.

The initial lexical translation sets undergo a first expansion step to capture morphological variation, using longest common prefix matching (hereafter, LCP). To apply prefix matching to the minimal set of elements necessary, we compute the following two set differences:

- Set of elements in the source to target translation set that are not members of the target token set: $T'_{ij} = T_{ij} - S_j$
- Set of elements in the target to source translation set that are not members of the source token set: $T'_{ji} = T_{ji} - S_i$

For each element in T'_{ij} (respectively T'_{ji}) and each element in S_j (respectively S_i), if a common prefix is found with a minimal length of more than n characters, the prefix is added to both translation sets.[2]

This simplified approach to stemming removes the need to rely on manually constructed endings lists to compute similarity or on a complete morphological analyser, which might not be available at all for under-resourced languages. It is also computationally more efficient as it exploits the nature of the alignment problem to reduce the search space: instead of matching each source and target word against every potential ending, with hundreds of possible endings in some languages, only the prefixes of word pairs within the subsets created through set difference need to be compared using LCP.

Another set expansion operation is defined to handle 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 a different domain. While creating the previously defined lexical translation sets from truecased sentences, capitalised tokens that are not found in the translation tables are added to the translation sets. Numbers are similarly handled and added to the expanded sets, as they can also act as alignment indicators, in particular when they denote dates.

These two expansions steps are essential to a successful use of Jaccard similarity for comparable sentence alignment. For instance, LCP gives a 2.9 points improvement in F1 measure on the initial Basque-Spanish test set described in Section 4.1, whereas the NE/Number expansion resulted in a 1.3 points gain; the two expansions combined gave a 4.3 points increase in terms of F1 measure. For the English-Bulgarian pair on the initial Wikipedia test set, the gains were 3.7, 2.6 and 5.5, respectively. Combining the two operations thus contributed to the improvements over the state of the art described in Section 4.3.

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. This notably eliminates the use of stop word lists from the computation of similarity.

Although it builds on fairly standard ideas, such as the use of GIZA tables or the Jaccard index, the approach is original in its conjoined use of these elements with surface-based information and simple set-theoretic operations to form a similarity assessment mechanism that proved efficient on comparable corpora, as shown in the next section.

4 Comparable sentence alignment

We performed a systematic comparison between different approaches to comparable sentence alignment on a variety of comparable corpora and language pairs. This section describes the components of the experimental setup.

4.1 Corpora

Three core sets of corpora were used in the evaluation, which we describe in turn. The selected test sets, all manually aligned, were used in different settings with gradual amounts of alignment noise added to the original sets. The goal of noisification is to assess the behavior of each approach in different scenarios and evaluate their ability to properly align data from ideal conditions to gradually noisier environments, the latter being a more realistic case when dealing with comparable corpora.

The first corpus consists in the public datasets created within the Accurat project.[3] The corpus covers 7 language pairs, each one composed of English and an under-resourced language. The datasets contain manually verified alignments that were created from news articles. We noisified these datasets by adding sentences from the

[2] Throughout the experiments we describe, n was set to 3.

[3] http://www.accurat-project.eu/. The corpus is available from: http://metashare.elda.org/repository/search/?q=accurat

TEST SETS	EN-DE	EN-EL	EN-ET	EN-LT	EN-LV	EN-RO	EN-SL
1:1	ATS: 512	ATS: 512	ATS: 512	ATS: 512	ATS: 512	ATS: 512	ATS: 512
2:1	ATS: 512 AOC: 512	ATS: 512 AOC: 512	ATS: 512 AOC: 512	ATS: 512 AOC: 512	ATS: 512 AOC: 512	ATS: 512 AOC: 512	ATS: 512 AOC: 512
100:1	ATS: 512 AOC: 6891 EUP: 43797	ATS: 512 AOC: 24276 EUP: 26412	ATS: 512 AOC: 50688	ATS: 512 AOC: 50688	ATS: 512 AOC: 50688	ATS: 512 AOC: 50688	ATS: 512 AOC: 15857 EUP: 34831

Table 1: Accurat evaluation sets

TEST SETS	BG-EN	DE-EN	ES-EN
1:1	WTS: 516	WTS: 314	WTS: 500
100:1	WTS: 516 EUP: 51084	WTS: 314 NC: 31086	WTS: 500 NC: 49500

Table 2: Wikipedia evaluation sets

TEST SETS	ES-EU
1:1	500-500
EITB_NOISE1	1000-1000
EITB_NOISE2	1000-1500

Table 3: EITB evaluation sets

original comparable corpora collected within the project, creating the following additional variants: (i) a 2:1 noisified version, where for each sentence in the original sets, 2 additional sentences without corresponding alignments were added; and (ii) a 100:1 noisified version with 100 sentences added for each sentence in the test sets. For each language pair, the additional sentences were taken from the initial portion of the selected additional corpora in one language and the final portion in the other language. For the 2:1 datasets, and the 100:1 variants in some language pairs, the original comparable corpora were used as additional data. For other language pairs, creating the 100:1 variant required adding sentences from different corpora to reach the required amount of data. Table 1 describes the final datasets used in the evaluation.[4]

As a second corpus, we used the data described in (Smith et al., 2010).[5] The texts were extracted from Wikipedia articles in 3 language pairs (English-German, English-Spanish and English-Bulgarian) and manually annotated for parallelism. We used the provided test sets (hereafter, WTS) and added a 100:1 noisified variant using sentences from the News Crawl corpus[6] for English-German and English-Spanish, and from Europarl for the English-Bulgarian pair. Table 2 describes these datasets, to which we will refer collectively as the Wikipedia corpus.

Finally, we used the EITB corpus, composed of news generated by the Basque Country's public broadcasting service.[7] The news are written independently in Basque and Spanish but refer to the same specific events and the corpus can thus be categorized as strongly comparable. We defined initial test sets of 500 manually aligned sentences in each language, and created two noisified variants: (i) a test set with 500 additional sentences in both languages, and (ii) a test set with 500 additional sentences in Spanish and 1000 in Basque. All additional sentences were taken from unaligned portions of the same EITB corpus. Table 3 summarises the EITB test sets.

The selected corpora thus cover 10 different language pairs and different domains, with varying degrees of noisification, and provide for a large and diverse comparison set.

4.2 Systems

Three approaches were evaluated against the previously described corpora: LEXACC (Stefănescu et al., 2012), the STACC method described in Section 3, and the approach based on Conditional Random Fields described in (Smith et al., 2010), to which we will refer as CRF. The latter was only evaluated on the Wikipedia corpus, using the re-

[4]In the table, ATS refers to the Accurat test sets, AOC to the Accurat original corpora, and EUP to the Europarl corpus.

[5]Available at: http://research.microsoft.com/en-us/people/chrisq/wikidownload.aspx.

[6]Refered to as NC here and available from: http://www.statmt.org/wmt13/translation-task.html.

[7]Euskal Irrati Telebista (EITB): http://www.eitb.eus. The corpus was provided courtesy of EITB and will be made available to the research community.

sults reported in the aforementioned article, as the tools to apply this method were not available to us; both LEXACC and STACC were evaluated on all test sets.

LEXACC was selected given its reported performance and its aim at portability across domains and language pairs; the system is also available as part of the Accurat toolkit,[8] which allowed for a direct comparison with STACC on all datasets.

The CRF approach has proven more effective than standard classifier-based methods on the Wikipedia datasets, with published results on publically available test sets, and was thus selected as an alternative approach to comparable sentence alignment.

Both approaches are based on sophisticated methods with demonstrated improvements over the state-of-the-art, thus providing strong baselines for system comparison.

4.2.1 LEXACC

LEXACC is a fast parallel sentence mining system based on a cross-linguistic information retrieval (CLIR) approach. It uses the Lucene search engine[9] in two major steps: target sentences are first indexed by the search engine, and a search query is built from a translation of content words in the source sentence to retrieve alignment candidates. The query is constructed using IBM Model 1 lexical translation tables, extracted from seed parallel corpora

The alignment metric in LEXACC is a translation similarity measure based on 5 feature functions briefly described here (see (Ștefănescu et al., 2012) for a detailed description):

- f_1 measures source-target candidate pairs strength in terms of content word translation and string similarity;
- f_2 is similar to f_1 but applies to functional words, as identified in manually created stop word lists;
- f_3 measures content word alignment obliqueness defined as a discounted correlation measure;
- f_4 is a binary feature that compares the number of initial/final aligned word translations over a pre-defined threshold;
- f_5 is a second binary feature which evaluates if the source and target sentences end with the same punctuation.

The similarity measure is then computed according to the sum of weighted feature functions, with optimal weights determined by means of logistic regression. We used the optimal feature weights described in (Ștefănescu et al., 2012) for the language pairs in the Accurat corpus and the provided default weights for English-Spanish and English-Bulgarian; for Basque-Spanish, optimal weights were estimated through logistic regression on a training set formed with 9500 positive parallel examples from the IVAP corpus[10] and an equal amount of non-parallel negative examples.

For the experiments, all lexical translation tables were created with GIZA++ on the JRC-Acquis Communautaire corpus.[11] Lucene searches were set to return a maximum of 100 candidates for each source sentence. We used the default setup for LEXACC, except for two minor changes. First, we removed the initial Lucene search constraint which was set to discard identical source and target sentences, a setting which prevented the retrieval of valid news candidates such as sports results. Secondly, we increased the length ratio filter from 1.5 to 7.5, as the initial value was too restrictive for the Basque-Spanish corpus. Both changes were thus meant to retrieve the most accurate set of alignment candidates, in order to get meaningful results on the test sets with both methods.

4.2.2 Conditional Random Fields

The model we refer to as CRF (Smith et al., 2010) is a first order linear chain Conditional Random Field (Lafferty et al., 2001), where for each source sentence a hidden variable indicates the corresponding target sentence to which it is aligned, or null if there is no such target sentence. This system was compared to the standard binary classifier of (Munteanu and Marcu, 2005) and to a ranking variant designed by the authors to avoid class imbalance issues that arise with binary classification. On the Wikipedia test sets, the CRF approach gave

[8] http://www.accurat-project.eu/index.php?p=accurat-toolkit

[9] http://lucenenet.apache.org/

[10] Extracted from the translation memories released by the Basque Public Administration Institute (http://opendata.euskadi.eus/catalogo/-/memorias-de-traduccion-del-servicio-oficial-de-traductores-del-ivap/), which consist of professional translations of public administration texts.

[11] We used the latest available version of the corpus, as of November 2015, in the OPUS repository: http://opus.lingfil.uu.se/JRC-Acquis.php.

the best results overall and was thus selected for our system comparison.

The sequence model comprises the following features:

- A word alignment feature set, based on IBM Model 1 and HMM alignments, which includes: log probability of the alignment; number of aligned/unaligned words; longest aligned/unaligned sequence of words; and number of words for different degrees of fertility.
- Two sentence-related features: source and target length ratio modeled through a Poisson distribution (Moore, 2002), and relative position of source and target sentences in the document.
- A set of distortion features measuring the difference in position between the previous and current aligned sentences.
- A set of features based on Wikipedia markup, including matching and non-matching links for alignment candidates.
- A set of lexicon features based on a probabilistic model of word pair alignments, trained on a set of annotated Wikipedia articles. The lexicon-based feature set includes the HMM translation probability, word-based positional differences, orthographic similarity, context translation similarity and distributional similarity.

The seed parallel data were based on the Europarl corpus for Spanish and German and the JRC-Aquis corpus for Bulgarian. The authors also included article titles of parallel Wikipedia documents and Wiktionary translations as additional seed data.

4.2.3 STACC

In order to establish a fair comparison between LEXACC and STACC, all shared settings were identical. Thus, lexical translations were based on the same previously described GIZA tables extracted from the JRC corpus, and STACC alignment was performed on the same sets of candidates retrieved from the Lucene searches by LEXACC for each language pair.

As described in Section 3, STACC is based on the k-best translations provided by lexical translation tables. For the experiments, k was set to 5, a value arbitrarily determined to be an optimal compromise between overcrowding the sets with unlikely translations and limiting translation candidates to minimal translation variants. Experimenting with different values on the test sets showed that this value for k was not actually the optimal one for some language pairs, with e.g. a 2.9 point gain in F1 measure when setting k to 2 for English-Greek on the initial Accurat test set.[12]

The results we present in the next section are thus not the best achievable ones using the STACC approach. Nonetheless, we maintained the use of a default value because of the lack of in-domain development sets on which an optimal value could be fairly computed.

4.3 Results

To evaluate the accuracy of the tested methods, precision was taken as the ratio of correct alignments over predicted alignments, and recall as the ratio of correct alignments over true alignments. We present results in terms of F1 measure, as we seek an optimal balance between alignment precision and recall.

Table 4 presents the results on the Accurat test sets for LEXACC and STACC using their respective optimal similarity thresholds.[13] On the 21 test sets, the two systems were tied on two occasions, with STACC obtaining better results in 89.5% of the remaining cases. On the noisiest datasets, STACC was consistently and markedly better across language pairs.

The results on the Wikipedia test sets are shown in Table 5. For English-Spanish and English-German, both approaches performed quite similarily on the initial test sets, with STACC obtaining the best results on the noisier sets.

The results for English-Bulgarian are interesting, as this is the only case where LEXACC outperforms STACC on both the clean and noisy datasets. The data used for noisification in this case may have had an effect on the results. Data extracted from Europarl, which compose the entire noisifi-

[12] Note that similar issues would arise if the selected translations were determined based on thresholds over translation probabilities, as the thresholds would need to be empirically set as well.

[13] The optimal thresholds were determined as the values providing the best results on the test sets. This would obviously not be an available threshold selection method when mining comparable corpora, where a default value would have to be used instead. Such a default value would however not allow for a fair comparison of the systems.

SYSTEM	TEST SETS	EN-DE	EN-EL	EN-ET	EN-LT	EN-LV	EN-RO	EN-SL
LEXACC	1:1	96.0	**89.5**	88.9	93.1	95.0	**99.4**	88.5
STACC	1:1	**96.7**	88.0	**92.0**	**96.1**	**96.6**	98.8	**89.5**
LEXACC	2:1	83.4	**83.2**	73.9	81.2	83.8	**95.3**	81.6
STACC	2:1	**89.2**	**83.2**	**79.9**	**86.9**	**88.2**	**95.3**	**82.3**
LEXACC	100:1	16.6	22.7	34.2	45.1	45.1	70.4	24.9
STACC	100:1	**33.7**	**37.3**	**42.5**	**56.0**	**56.2**	**75.7**	**35.3**

Table 4: Best F1 measures on the Accurat evaluation sets

SYSTEM	TEST SETS	EN-BG	EN-DE	EN-ES
LEXACC	1:1	**87.1**	**82.7**	98.2
STACC	1:1	84.9	82.0	**99.7**
LEXACC	100:1	**27.6**	31.0	66.2
STACC	100:1	16.6	**35.8**	**73.3**

Table 5: Best F1 measures on the Wikipedia evaluation sets

	CRF		LEXACC		STACC	
LANGUAGE PAIR	R@90	R@80	R@90	R@80	R@90	R@80
EN-BG	72.0	81.8	**80.4**↑	80.4↑	80.2	**81.6**↑
EN-DE	58.7	68.8	**75.2**	78.7	68.8	**81.8**↑
EN-ES	90.4	93.7	97.0↑	97.0↑	**99.6**↑	**99.6**↑

Table 6: Targeted recall on the Wikipedia evaluation sets

SYSTEM	TEST SETS	ES-EU
LEXACC	1:1	77.2
LEXACC_DF	1:1	80.2
STACC	1:1	**90.9**
LEXACC	EITB_NOISE1	59.2
LEXACC_DF	EITB_NOISE1	62.2
STACC	EITB_NOISE1	**82.8**
LEXACC	EITB_NOISE2	54.5
LEXACC_DF	EITB_NOISE2	57.4
STACC	EITB_NOISE2	**79.5**

Table 7: Best F1 measures on the EITB evaluation sets

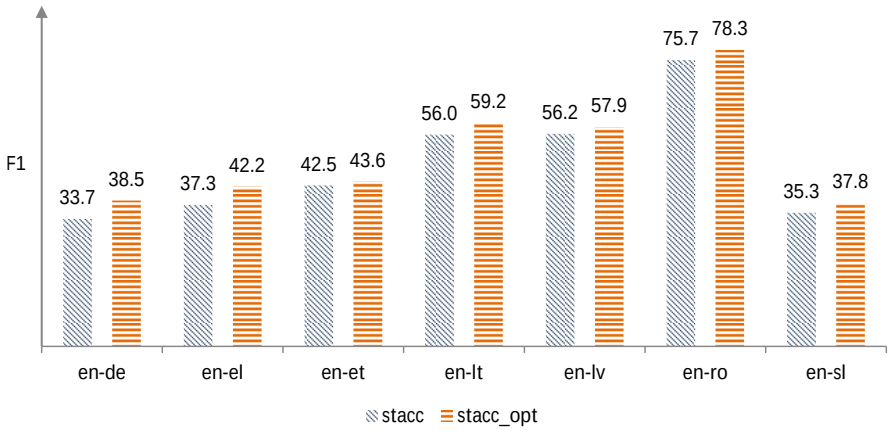

Figure 1: STACC optimisation results on the Accurat 100:1 test sets

cation set for this language pair, is closer to the JRC vocabulary than the original comparable data on which the alignment process would take place in real-world conditions. Although we have not thoroughly tested the impact of this variable, it is possible that those datasets are more confusing for an approach such as STACC, which is based mostly on lexical information extracted from seed parallel data, than for a feature-based approach where some features, like the boolean punctuation-based ones in LEXACC, may compensate for erroneous alignments due to artificial domain vocabulary overlap. Determining if this hypothesis is indeed correct would require further experiments beyond the scope of this paper

To include the CRF approach in the comparison, we used two of the provided measures, namely recall obtained at precisions of 80 and 90 percent on the 1:1 test sets.[14] We report results obtained with the best variant of CRF, namely the model which includes Wikipedia and lexicon features, with intersected results from both directions. Results are reported in Table 6. Although the comparison was limited in this case, results were in favour of LEXACC and STACC on targeted recall measures for the Wikipedia datasets.

Finally, both LEXACC and STACC were compared against the EITB test sets, with results shown in Table 7. For this language pair, STACC performed markedly better with differences of up to 25 points. A likely explanation for these results is the nature of the features that compose the LEXACC model. In particular the features related to alignment obliqueness and number of initial/final aligned words might be detrimental in the case of Basque, which exhibits free word order. Given the poor results obtained with feature weights optimised on the IVAP corpus, we also checked the results using the provided default weights. This resulted in slightly better performance, as shown in the rows named LEXACC_DF in Table 7, though still far from the results achieved with STACC.

4.4 Discussion

Overall, STACC provided the best results across domains and language pairs, in particular for noisier datasets. Additionally, the approach has several advantages over existing methods and systems for comparable segment alignment.

First, it is undoubtedly simpler, as it requires but minimal information to reach optimal results. Lexical tables and simple set expansion operations based on surface properties of the tokens are the only components of the approach, as compared to the more sophisticated feature-based approaches which rely on larger sets of components for which optimal weights need to be computed prior to applying the models.

Secondly, because of its simplicity, STACC is a more portable method, as is it is not necessary to perform any type of adaptation for new domains and language pairs, nor to rely on domain-specific information such as link structure in Wikipedia. In actual practice, portability is an important issue which hinders on the exploitation of comparable corpora. An efficient yet easily deployable method is therefore a welcome addition to the toolset for parallel data extraction.

Finally, STACC results in fewer computational steps when compared to more complex feature-based methods. First, it involves simple binary set intersection and union operations for the computation of similarity, instead of conjoined feature computation on larger component sets. Secondly, the approach relies on tractable set differences for its most computationally expensive operation of longest common prefix matching, compared to matching all tokens against lists of word endings which can be quite large, notably in the case of agglutinative languages.

Although promising, the approach could be further evaluated, and potentially improved, along two main lines.

It might be worth exploring for instance the impact of filtering alignment candidates according to the relative position of sentence pairs in the original source and target documents, a document-level property notably exploited by (Smith et al., 2010). As the STACC approach is featureless, and meant to remain as such in order to maintain its portability and ease of deployment, filtering distant sentence pairs would need to take place prior to the computation of alignment scores. A simple approach compatible with STACC would consist in constraining candidate sets by including sentence position information when performing indexing and candidate querying in a CLIR approach. This would provide an additional evalua-

[14] Note that, for both LEXACC and STACC, in some scenarios even the lowest thresholds gave precisions higher than 90, rendering the comparison moot. We indicate these cases with a ↑ sign next to the highest recall obtained at the closest precision to the arbitrary 80 and 90 precision points.

tion of the accuracy of the approach in scenarios where document-level information is exploitable.

Additionally, given the importance of k-best lexical translations in computing STACC similarity, variations in lexical coverage obtained with different translation tables can be expected to impact alignment accuracy. Although mining comparable corpora usually requires the use of seed translation knowledge extracted from a domain that differs from the one being mined, default tables with wide lexical coverage can be built from existing parallel corpora in different domains. Thus, improvements might be obtained with larger and more diverse tables than the ones used in the experiments reported here, which were based on translations extracted from a single domain. A precise assessment of the evolution of alignment accuracy given variations in lexical translation coverage is left for future research.

5 Alignment optimisation

As previously mentioned, for both LEXACC and STACC, alignments were computed for every source sentence against candidate translations retrieved by Lucene and all cases where a given target sentence has more than one source alignment were left as is.

Although this methodology enabled a fair comparison between the two systems, it evidently impacts alignment accuracy. One simple optimisation is to retain only the best overall source-target alignments, discarding all alignments established between a given source sentence and a target sentence if the latter is linked to better scoring source sentences.

The net effect of this procedure is the promotion of better alignments, as some correct alignments would not be hidden anymore by other better scoring shared alignments. This is most likely to occur with source-target pairs that are close variants of each other, with close similarity scores.

We applied this simple optimisation to the Accurat test sets and observed improvements across the board, as shown in Figure 1. Depending on actual usage, this optimised version of STACC alignment can constitute the best alternative for the extraction of parallel sentences from comparable corpora.

6 Conclusions

We described a simple approach to comparable sentence alignment, termed STACC, which is based on automatically extracted seed lexical translations, the Jaccard similarity coefficient, and simple set expansion operations that target named entities, numbers, and morphological variation using longest common prefixes. Building on fairly standard components for the computation of similarity, this method is shown to perform better than current alternatives.

The approach was evaluated on a large range of datasets from various domains for ten language pairs, giving the best results overall when compared to sophisticated state-of-the-art methods. STACC also performed better than competing approaches on noisier corpora, showing promises for the exploitation of the typically noisy data found when mining comparable corpora.

STACC is a highly portable method which requires no adaptation for its application to new domains and language pairs. It thus allows for the fast deployment of a crucial component in comparable corpora alignment, which opens the path for an increase in the amount of such corpora that can be exploited in the future.

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), PLATA (IG-2014/00037) and TRADIN (IG-2015/0000347) projects. We would like to thank MondragonLingua Translation & Communication as coordinator of these projects and the three 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*, EACL '09, pages 16–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

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.

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.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

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*, pages 289–295. Association for Computational Linguistics.

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

Makoto Nagao. 1984. A Framework for a Mechanical Translation Between Japanese and English by Analogy Principle. In *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, New York, NY, USA. Elsevier North-Holland, Inc.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Monica Lestari Paramita, David Guthrie, Evangelos Kanoulas, Rob Gaizauskas, Paul Clough, and Mark Sanderson. 2013. Methods for collection and evaluation of comparable documents. In *Building and Using Comparable Corpora*, pages 93–112. Springer.

Mohammad Taher Pilehvar, David Jurgens, and Roberto Navigli. 2013. Align, disambiguate and walk: A unified approach for measuring semantic similarity. In *Proceedings of the 51st meeting of the Association for Computational Linguistics*, pages 1341–1351. The Association for Computational Linguistics.

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.

Inguna Skadiņa, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su, Dan Tufis, Mateja Verlic, Andrejs Vasiļjevs, Bogdan Babych, Paul Clough, Robert Gaizauskas, et al. 2012. Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*.

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*, HLT '10, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Dan Ştefănescu, 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*, pages 745–748. IEEE.