# Exploiting a Large Strongly Comparable Corpus

**Thierry Etchegoyhen, Andoni Azpeitia and Naiara Pérez**
Vicomtech-IK4
Donostia / San Sebastián, Spain
{tetchegoyhen, aazpeitia, nperez}@vicomtech.org

**Abstract**

This article describes a large comparable corpus for Basque and Spanish and the methods employed to build a parallel resource from the original data. The EITB corpus, a strongly comparable corpus in the news domain, is to be shared with the research community, as an aid for the development and testing of methods in comparable corpora exploitation, and as basis for the improvement of data-driven machine translation systems for this language pair. Competing approaches were explored for the alignment of comparable segments in the corpus, resulting in the design of a simple method which outperformed a state-of-the-art method on the corpus test sets. The method we present is highly portable, computationally efficient, and significantly reduces deployment work, a welcome result for the exploitation of comparable corpora.

**Keywords:** Comparable corpora, Alignment, Corpus exploitation, Basque, Statistical machine translation

## 1. Introduction

Comparable corpora are a useful resource to overcome the issue of scarce parallel data in some language pairs and domains, providing statistical machine translation systems (Brown et al., 1990; Koehn, 2010) with the necessary data to increase overall MT quality (Munteanu and Marcu, 2005; Irvine and Callison-Burch, 2013). In this paper, we describe a large comparable corpus for Basque and Spanish and the methods employed to build a parallel resource from the original data.

Our first contribution is thus a description of the first large corpus for the Basque-Spanish language pair in the news domain, based on content professionally created by the EITB Basque public broadcasting services. This resource will be shared with the research community, as a basis for further advances in comparable corpora exploitation and data-driven machine translation. Additionally, the resource will contribute to advancing research on a minority language with scarce publically available resources.

The second contribution of this paper is the evaluation of a simple method for the alignment of comparable segments,[1] which improves significantly over the state-of-the-art aproaches we tested on the corpus in terms of F1 measure. The approach we describe, which is based on the Jaccard index and lexical set expansion, also improves over existing methods in terms of simplicity and deployment effort.

We first describe the EITB and IVAP corpora in Section 2, which feed the alignment methods described in the remainder of the paper. Section 3 presents the alignment methods we used for document and segment aligment. In section 4, we discuss the results obtained on the EITB corpus. Finally, Section 5 presents concluding remarks and a description of future work.

[1]Hereafter, we use the term *segment* to refer to sentences as well as smaller independent linguistic units.

## 2. Corpora

The EITB corpus is composed of news written by journalists of the Basque Country's public broadcast service.[2] The news are written independently in Basque and Spanish but refer to the same specific events. The corpus can thus be categorized as strongly comparable, following standard terminology (Skadiņa et al., 2012), and its exploitation should provide a solid basis for the development of generic Basque-Spanish SMT systems.

The original dataset is composed of 59 XML documents in Basque and 57 in Spanish, covering five years of news content generation, from 2009 to 2013. Each one of these files contains varying amounts of news items with the following structure:

- <**id**>: An integer identifying the news item. Note that these numbers have no established correspondence between languages, i.e. they cannot be used to align news between languages.
- <**title**>: The title of the news item.
- <**link**>: The original HTML publication link.
- <**pubDate**>: Date of publication.
- <**description**>: Textual content of the news item, typically amounting to one or two small paragraphs.
- <**category**>: Indicates the broad category to which the news item belongs (Culture, Sports, Economy, etc.).

As specific news are generated independently, they can be considered independent documents and were extracted as such. Details about the corpus are described in Table 1.

With over one million sentences per language, this bilingual corpus is the largest available for the Basque-Spanish language pair. The corpus covers political news, sports and cultural events, among others, and thus offers a relatively broad representation in terms of topics and vocabulary.

[2]Euskal Irrati Telebista: http://www.eitb.eus.

| Year | EU News | ES News | EU Sentences | ES Sentences | EU Tokens | ES Tokens |
|---|---|---|---|---|---|---|
| 2009 | 18,552 | 18,759 | 236,753 | 223,323 | 3,068,989 | 4,672,018 |
| 2010 | 17,462 | 17,979 | 216,043 | 204,004 | 2,778,677 | 4,325,927 |
| 2011 | 18,856 | 19,037 | 230,902 | 216,240 | 3,083,384 | 4,948,890 |
| 2012 | 19,344 | 18,972 | 229,270 | 213,730 | 3,043,726 | 4,932,887 |
| 2013 | 13,484 | 13,601 | 164,363 | 160,908 | 2,160,011 | 3,557,014 |
| **Total** | **87.698** | **88.348** | **1.077.331** | **1.018.205** | **14.134.787** | **22.436.736** |

Table 1: Original EITB corpus

To be able to align segments in the EITB comparable corpus, a minimal amount of bilingual correspondences is needed. At present, the only large bilingual dataset freely available for the Basque-Spanish pair is the collection of translation memories released by the Instituto Vasco de Administración Pública (IVAP),[3] which consist of professional translations of public administration texts.

We extracted the textual content in the original TMX translation memories, and performed sentence aligment with the HunAlign toolkit (Varga et al., 2005). After corpus cleanup, which involved removing segments with corrupt characters and erroneously tagged translation units, we prepared the parallel datasets shown in Table 2. Each sentence in the corpus was tokenized and truecased, using truecasing models trained on the original texts in both languages. The *train*, *dev* and *test* datasets were used to respectively train, tune and evaluate a phrase-based SMT system (Koehn et al., 2003), using the Moses toolkit (Koehn et al., 2007), which serves as a component for one of the alignment methods described in the next section. Lexical translation tables were created with the GIZA++ toolkit (Och and Ney, 2003).

## 3. Alignment

To extract a bitext from the original XML documents that constitute the EITB corpus, the source and target sets need to be mined for alignable segment pairs. An exhaustive search would apply to the Cartesian product of the source and target sets (Ion, 2012), a highly resource-consuming process. Alternative approaches that reduce the search space have been designed using document alignment as a first step (Fung and Cheung, 2004; Ion et al., 2011), or cross-language information retrieval (CLIR) techniques, where target segments are retrieved through search engine indexing and querying (Rauf and Schwenk, 2011; Munteanu and Marcu, 2005; Stefănescu et al., 2012). We first applied search space reduction through document alignment, and then performed segment alignment.

### 3.1. Document alignment

As news may have been generated in one language but not in the other, there is no strict one-to-one correspondence between the files that form the original document set. In order to reduce the search space of alignment candidates, we first performed document alignment using the EMACC tool (Ion et al., 2011),[4] which follows an Expectation Maximization approach for the alignment of textual units in comparable corpora.

The alignment was created with default parameter values, with a maximum of 3 target document alignments to consider. Given the size of the initial corpus and manual examination of the results, the default values offered a good compromise in terms of number of alignments and alignment quality.

The document alignment process with EMACC generated alignments for 94% of the Spanish source documents, and 92% of the target Basque documents —a satisfactory portion of the initial data considering that the alignment of comparable documents is a difficult task for which alignments can only be approximated.

### 3.2. Segment alignment

The data in the aligned documents described in the previous section served as input to the segment alignment methods. In order to test the accuracy of said methods, a dataset of 500 source and target sentences was manually aligned and verified. To test recall we built two additional datasets that extended this test set with unaligned data: a first dataset with 500 additional unaligned source and target sentences, and a second dataset with 500 additional unaligned source sentences and 1000 additional unaligned target sentences.

In the next sections, we describe the methods that were used for segment alignment and their results on these test sets.[5]

#### 3.2.1. LEXACC alignment

LEXACC (Stefănescu et al., 2012) is a fast parallel sentence mining algorithm, based on CLIR, which has proved effective for the task. It uses the Lucene search engine[6] 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 approximated translation used for the query is constructed using IBM model 1 lexical translation tables (Brown et al., 1993), extracted from seed parallel corpora using the GIZA++ toolkit. For our purposes, the tables were based on the IVAP corpus described in Section 2. LEXACC is part of the Accurat toolkit and can be used in conjunction with EMACC, having Lucene searches narrowed down

[3] http://opendata.euskadi.eus/catalogo/-/memorias-de-traduccion-del-servicio-oficial-de-traductores-del-ivap/

[4] Available as part of the Accurat toolkit: http://www.accurat-project.eu/index.php?p=accurat-toolkit, see (Skadiņa et al., 2012).

[5] To keep the comparative segment alignment results as fair as possible, all preliminary steps before segment alignment (i.e., document alignment, indexing, queries, searches and results) have been kept strictly identical between methods. Sentences were likewise tokenised and truecased with identical tools and models, using the tools provided in the Moses distribution.

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

| Dataset | Aligned Sentences | EU Tokens | ES Tokens |
|---|---|---|---|
| train | 645,223 | 7,556,964 | 9,717,604 |
| dev | 2,000 | 37,908 | 48,492 |
| test | 2,000 | 38,081 | 49,056 |

Table 2: IVAP corpus

to the documents aligned by the document aligner. We followed this approach for the experiments presented here, using the document alignments described in Section 3.1.
The alignment metric in LEXACC is a translation similarity measure based on 5 feature functions briefly described here (see (Stefănescu et al., 2012) for a detailed description):

- $f_1$: A feature measuring the source-target candidate pair's strength in terms of content word translation, based on named entity matching, string similarity and lexical translation probability scores as given by GIZA tables.
- $f_2$: A feature similar to $f_1$ but applying to functional words, as identified in manually created stop word lists.
- $f_3$: This feature measures alignment obliqueness (Tufiş et al., 2006), with crossing content word alignments viewed as an estimator of source-target alignment strength.
- $f_4$: A binary feature modeling the assumed tendency of parallel segments to start, respectively end, with aligned word translations.
- $f_5$: A second binary feature with value 1 if both the source and target segments end with the same punctuation, and 0 otherwise.

The similarity measure is then computed according to the sum of weighted feature functions, with optimal weights determined by means of logistic regression. The optimized weights were computed on a training set formed with 9500 positive parallel examples from IVAP and an equal amount of non-parallel negative examples; the evaluation set contained 500 positive and 500 negative examples. Final scores are computed in both translation directions and symmetrized by arithmetic average.
An accuracy of $0.95$ was obtained on the IVAP test sets with the optimized weights; results on the EITB test sets are presented in Section 4.

#### 3.2.2. Feature adaptation

Given the morpho-syntactic properties of Basque and initial results, we explored variants of the original feature set in LEXACC. Features $f_3$ and $f_4$ were prime candidates for replacement, as free word order in Basque and mirror syntactic structures in both languages casted doubt on the usefulness of features based on crossing alignments or finding corresponding alignments in specific sentential positions. Furthermore, as several classes of functional words are agglutinated in Basque, and proper segmentation would require the use of a complete morpho-syntactic analyser, feature $f_2$ was considered as a potential candidate for replacement as well. We experimented with several combinations of features, and selected the following four:

- $f'_1$: A feature measuring the source to target candidate pair's strength in terms of content word translation.
- $f'_2$: A feature measuring the target to source candidate pair's strength in terms of content word translation.
- $f'_3$: A feature measuring the amount of common entities between source and target segments, where entities were defined as capitalized in-sentence tokens not found in the lexical translation tables.
- $f'_4$: The original binary feature $f_5$, which indicates if the source and target segments end with the same punctuation.

Additionally, we experimented with simplified similarity scoring: for the first 3 features, the score was incremented by 1 if a corresponding word was found in either the lexical tables or as an entity, instead of using translation probabilities directly. The reason for this change was the difference in domain between the IVAP and EITB corpora: lexical translation probabilities established for one domain were not necessarily representative of lexical distributions in the other domain, and entities were likely not to overlap between the two domains. In Section 4 we show that this adapted feature set and scoring gave significant improvements over the original LEXACC setup.

#### 3.2.3. STACC alignment

Feature-based approaches to segment similarity can raise difficulties in terms of adaptation to new language pairs. Although the original LEXACC feature set was designed for cross-linguistic application, the features were not optimal for the Basque-Spanish pair and additional work was necessary to determine and test better feature sets, along with scoring variants. This type of time-consuming adaptation effort seems unavoidable in any feature-based approach to segment similarity, in order to better exploit comparable corpora for new language pairs and domains.
Another issue comes from the computation of the similarity measure between segments. In LEXACC, a core part of the final score is computed by evaluating each source token against each target token and measuring translation correspondences using lexical translation probabilities for each token pair. This approach suffers from being a translation heuristic, which does not properly tackle translation fertility, nor the validity of lexical translation probabilities without contextual disambiguation.
Finally, the use of lists of stop words and endings is error-prone and can be computationally expensive for languages such as Basque with cascading agglutinative markers. Although these lists still need to be part of the Lucene query

building process, removing their use from the computation of segment similarity would be an improvement in terms of precision and efficiency.

To tackle these shortcomings, we explored an alternative approach, termed STACC,[7] whose main steps are described below. We describe two variants of the approach, one using a complete SMT system and the other relying on lexical translation tables only. Both variants were tested against the EITB test sets, with results presented in section 4.

The STACC approach relies on the Jaccard coefficient (Jaccard, 1901), which measures set similarity by taking the ratio of set intersection over union. This index has been standardly used as a measure of text similarity and applied to comparable documents by (Skadiņa et al., 2012) and (Paramita et al., 2013) for instance, where Jaccard similarity is computed between a portion of the sentences found in bilingual documents and overall document comparability is measured as the average of these sentence-level scores. We adopt and extend the use of this index in the STACC approach.

Let $s$ be the source segment, $C$ the set of target alignment candidates retrieved from a CLIR search-based step, as in the LEXACC approach, $T_s$ the set of translated source tokens and $K_c$ the set of tokens from a given candidate segment $c$ in $C$. The basic STACC similarity score is given in equation 1:

$$score_{stacc} = \frac{|T_s \cap K_c|}{|T_s \cup K_c|} \quad (1)$$

To account for morphological variation, we integrated a set expansion step using longest common prefixes (LCP): for each token pair in $s_t$ and current candidate $c_i$, if a common prefix is found with a minimal length of 3 characters, the prefix is added to both the $T_s$ and $K_c$ sets. This simplified approach to stemming removes the need to rely on manually constructed endings lists to compute similarity or on a complete morphological analyzer, which might not be available at all for under-resourced languages. It is also computationally more efficient: instead of matching each source and target word against every potential ending, with hundreds of possible endings in a language like Basque, only the prefixes of the two words in the considered pairs need to be compared using LCP. Furthermore, the operation is performed on the set differences between $T_s$ and $K_c$, i.e. to the tokens that are not members of the other set, thus applying to the minimal relevant sets of tokens.

The second set expansion operation is meant to account for potential named entities and involves including in both set representations all tokens that are present in capitalised form; numbers are added as well, as they can act as indicators of comparability, when they represent dates in particular, and are likely to be absent from translation tables. This operation is performed while creating the set representations from the initial tokenised and truecased strings, thus adding negligible computational cost.

Finally, note that no filtering is performed, with punctuation and functional words kept alongside content words in the final sets.

[7]Which stands for *Set-Theoretic Alignment for Comparable Corpora*.

On the basis of this core method involving the Jaccard index and set expansions, we tested a first variant, termed STACC_MT, where the initial set of translated tokens $T_s$ is created by machine translating the source segment using a phrase-based SMT system trained on the IVAP parallel corpus, as mentioned in section 2. As lexical translation tables are a prerequisite for any type of comparable segment alignment, a core set of parallel segments is needed in any case, from which a phrase-based SMT system can be built without additional resources. This approach relies on the precision of the complete SMT system to provide the optimal translation for the source sentence, handling translation fertility and phrasal translation through the tuned log-linear combination of features that define the model.[8] The score in this variant is computed according to equation 1.

The second variant, to which we will refer as STACC_LEX, relies only on lexical translation tables. Two translated sets are used to compute the similarity score: $T_{st}$ contains the k-best lexical translations for each source token, and $T_{ts}$ the k-best lexical translations for each token in the alignment candidate.[9] With $S$ as the set of tokens in the source segment $s$, the Jaccard scores are computed in each direction and averaged, as shown in Equation 2.

$$score_{stacc\_lex} = \frac{\frac{|T_{st} \cap K_c|}{|T_{st} \cup K_c|} + \frac{|T_{ts} \cap S|}{|T_{ts} \cup S|}}{2} \quad (2)$$

This variant has the potential advantage of capturing more translation options than the first approach based on machine-translated candidates, through the use of sets of lexical translations. It also enables the use of bi-directional translation information without the need to train a complete MT system in both directions. Finally, it removes the need to fully translate the source segments, thus providing for a faster computation of the overall alignment. Note that the core of this approach is similar to the symmetrised use of lexical translations in LEXACC as a component feature in their model, with the marked difference that the actual probabilities are discarded and replaced with set membership in order to minimise the impact of domain differences between lexical tables and the comparable corpora at hand.

The STACC approach thus addresses the identified shortcomings and offers a simple, principled and more efficient computation of alignments. These advantages translate into better alignment results on our reference sets, and better material for the exploitation of the corpus, as shown in the following section.

## 4. Results

Two sets of results are described below. We first present the alignment results that were obtained with the competing methods described in the previous section, then turn to an evaluation of the SMT systems that were trained on the different alignment sets.

[8]The use of fully machine translated source sentences as a basis to determine segment similarity has been explored by others, although with different similarity metrics. See for instance (Sarikaya et al., 2009), (Abdul-Rauf and Schwenk, 2009) and (Yasuda and Sumita, 2008).

[9]In the experiments we present, $k$ was set to 5.

### 4.1. Alignment results

Table 3 presents the best F1-measure scores achieved by each method, the similarity threshold at which it was achieved and the corresponding precision and recall scores. Precision was computed as the ratio of correct alignments over predicted alignments, while recall was measured as the ratio of correct alignments over total correct alignments.

It is worth noting that recall and precision were measured over the alignments assigned by each method on a single pass. Further processing can improve over these alignments by discarting, for each aligned source segment, those alignments for which a different source segment obtained a better score, with more correct alignments surfacing as a result. We do not present results along these lines to maintain a fair comparison between the LEXACC and STACC approaches, as the former does not include this additional processing step.

Since an appropriate goal for the extraction of parallel data from comparable corpora is to find the largest amount of alignments with satisfactory precision, the F1 measure is a good indicator of the fruitfulness of a given alignment method. On all three test sets, STACC alignment provides markedly better F1 scores in both variants of the approach, with the lexical variant giving the best results.

Interestingly, the feature-adapted version of LEXACC provides competitive results. One possible reason is the removal of features that were assumed to be detrimental in the case of Basque, as hypothesised in Section 3.2.2. A second likely explanation is the modified scoring that was employed in this version for the lexical translation feature, with translation probabilities discarded and a score of 1 being assigned to matching lexical translations. In effect, this renders the modified LEXACC score similar to set membership, thus approximating the results obtained with STACC.

Another interesting aspect of the comparative analysis we performed concerns the evolution of precision and recall given similarity scores. Figure 1 shows results along these lines, with a visualisation of average precision and recall scores for the three test sets.

Both methods exhibit similar evolutions in terms of recall and precision, the main difference being the comparatively smoother drop of LEXACC recall over a larger portion of the scoring range. This similarity demonstrates that the differences in terms of best F1 scores are not due to local scoring peaks but to the consistently higher marks obtained with STACC.

Overall, STACC provides the best balance between the number of alignments it can identify and the precision in identifying correct alignments. Thus, this method proved to be the optimal choice to retrieve the largest amount of precise alignments in the EITB corpus. On the entire EITB corpus, 368184 parallel segments were extracted using LEXACC and 596492 using the STACC_LEX variant; segment similarity thresholds were set at 0.33 for LEXACC and 0.30 for STACC, the optimal marks according to the results on the noisiest test sets. Details of the aligned corpus are shown in Table 4.

| Method | Segments | EU Tokens | ES Tokens |
|---|---|---|---|
| LEXACC | 368,184 | 5,756,838 | 7,605,144 |
| STACC | 596,492 | 10,875,355 | 16,537,244 |

Table 4: Extracted alignments

### 4.2. Translation results

One major goal in exploiting comparable corpora is the creation of parallel resources to help develop data-driven translation systems. In order to evaluate the impact of the different aligned datasets created by the methods we examined, we trained SMT systems and measured their quality in terms of the BLEU automated metric (Papineni et al., 2002).
All systems are phrase-based models built with the Moses toolkit, with phrases of length 5, a distortion limit of 6 and surface factors only. The language models were trained on the target side of the bitexts, using the KENLM toolkit (Heafield, 2011) to generate 5-gram models with modified Kneser-Ney smoothing (Heafield et al., 2013). Parameters of the log-linear models were tuned with MERT (Och, 2003) on a set of 2000 sentences extracted from the aligned sets and the models were evaluated using the Multeval toolkit (Clark et al., 2011) on test sets composed of 1678 manually verified aligned sentences.

| MT MODEL | EU → ES | ES → EU |
|---|---|---|
| IVAP | 13.7 | 9.1 |
| LEXACC | 23.7 | 17.8 |
| STACC | **25.5** | **19.1** |

Table 5: BLEU results for Basque-Spanish

Table 5 presents the results in terms of BLEU, taking the IVAP out-of-domain models as baselines. These results first show the gain obtained with the extracted in-domain EITB parallel corpora over the IVAP baselines, providing another example of the value of comparable corpora for the development of SMT models. The results also show the significant improvements obtained with STACC alignment, confirming the value of the approach.

## 5. Conclusion

In this paper, we described the EITB corpus, a large strongly comparable Basque-Spanish corpus in the news domain. This resource is to be shared with the research community, as an aid for the development and testing of methods in comparable corpora exploitation, and as a basis for the improvement of data-driven machine translation systems for this language pair.[10]
We also explored several methods for the alignment of comparable segments in the corpus, using publically available resources and assessing their usefulness for the corpus at hand. This work resulted in the design of the STACC similarity metric, based on the Jaccard index and expanded lex-

[10]The corpus will be made available in the META-SHARE repository (http://www.meta-share.eu/) as the $Basque\_Spanish\_EITB\_comparable\_corpus$, under the CC - BY - NC - SA licence for academic users and the MS - C - NO RED - FF licence for commercial users.

| Method | Testset | Threshold | Precision | Recall | F1 |
|---|---|---|---|---|---|
| LEXACC | EITB | 0.12 | 80.5 | 74.2 | 77.2 |
| FEAT_ADAPT | EITB | 0.05 | 85.4 | 85.4 | 85.4 |
| STACC_MT | EITB | 0.15 | 88.6 | 87.0 | 87.8 |
| **STACC_LEX** | EITB | 0.17 | **91.0** | **90.8** | **90.9** |
| LEXACC | EITB *Noise 500+500* | 0.31 | 63.3 | 55.6 | 59.2 |
| FEAT_ADAPT | EITB *Noise 500+500* | 0.36 | 76.3 | 69.0 | 72.5 |
| STACC_MT | EITB *Noise 500+500* | 0.24 | 80.2 | 70.6 | 75.1 |
| **STACC_LEX** | EITB *Noise 500+500* | 0.30 | **84.4** | **81.2** | **82.8** |
| LEXACC | EITB *Noise 500+1000* | 0.33 | 59.6 | 50.2 | 54.5 |
| FEAT_ADAPT | EITB *Noise 500+1000* | 0.36 | 72.2 | 66.6 | 69.3 |
| STACC_MT | EITB *Noise 500+1000* | 0.24 | 79.0 | 67.6 | 72.8 |
| **STACC_LEX** | EITB *Noise 500+1000* | 0.30 | **81.1** | **78.0** | **79.5** |

Table 3: Best F1 measures

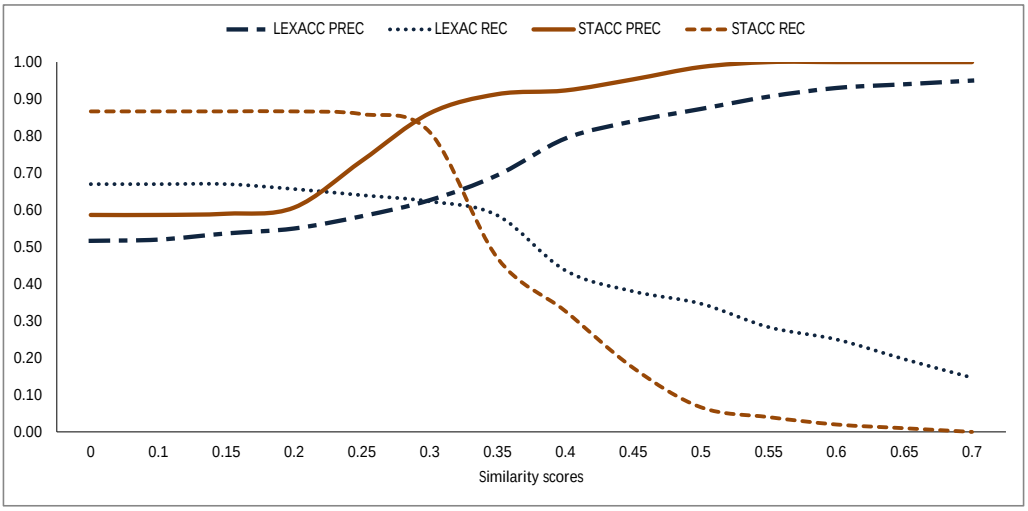


Figure 1: Evolution of average precision and recall

ical sets. We experimented with two variants of the approach, both of which outperformed a state-of-the-art approach on our test sets. This simple approach gave optimal results in terms of segment alignment, as well as BLEU scores of the derived SMT systems.

The method we presented is highly portable, computationally efficient, and significantly reduces deployment work, a welcome result for the exploitation of comparable corpora. In future work, we will explore its usefulness in other domains and for different language pairs.

**Acknowledgements** The authors wish to thank Euskal Irrati Telebista, for providing the resources and agreeing to share them with the research community, and the three anonymous LREC reviewers for their constructive feedback. This work was partially supported by the Basque Government through its funding of project PLATA (Gaitek Programme, 2012-2014).

## 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.

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.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Barcelona, Spain*, pages 57–63.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified kneser-ney language model

estimation. In *ACL (2)*, pages 690–696. The Association for Computer Linguistics.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ion, R., Ceauşu, A., and Irimia, E. (2011). An expectation maximization algorithm for textual unit alignment. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 128–135. Association for Computational Linguistics.

Ion, R. (2012). PEXACC: A parallel sentence mining algorithm from comparable corpora. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey*, pages 2181–2188.

Irvine, A. and Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270.

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.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics-Volume 1*, pages 48–54. Association for Computational Linguistics.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180. Association for Computational Linguistics.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press.

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.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Paramita, M. L., Guthrie, D., Kanoulas, E., Gaizauskas, R., Clough, P., and Sanderson, M. (2013). Methods for collection and evaluation of comparable documents. In *Building and Using Comparable Corpora*, pages 93–112. Springer.

Rauf, S. A. and Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine translation*, 25(4):341–375.

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, Brighton, UK*, pages 432–435.

Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., et al. (2012). Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey*.

Stefănescu, 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.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Yasuda, K. and Sumita, E. (2008). Method for building sentence-aligned corpus from wikipedia. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence*.