Making the most of comparable corpora in Neural Machine Translation: a case study

Harritxu Gete^{1,2} · Thierry Etchegoyhen¹

Accepted: 10 December 2021 / Published online: 27 January 2022
© The Author(s) 2021

Abstract Comparable corpora can benefit the development of Neural Machine Translation models, in particular for under-resourced languages. We present a case study centred on the exploitation of a large comparable corpus for Basque-Spanish, created from independently-produced news by the Basque public broadcaster EITB, where we evaluate the impact of different techniques to exploit the original data, in order to complement parallel datasets for this language pair in both translation directions. Two efficient methods for parallel sentence mining are explored, which identified a common core of approximately half of the total number of aligned sentences, each one uniquely identifying valid parallel sentences not captured by the other method. Filtering the data via identification of length-difference outliers proved highly effective to improve the models, as was the use of tags to discriminate between comparable and parallel data in the training corpora. The use of back-translated data is also evaluated in this work, with results indicating that alignment-based datasets remain the most beneficial, although complementary backtranslations should also be included to fully exploit the available comparable data. Overall, the results in this work demonstrate that this type of data needs to be carefully analysed prior to its use as training data for Neural Machine Translation, since issues such as information imbalance between source and target data can lead to unoptimal results for a given translation pair.

Harritxu Gete and Thierry Etchegoyhen have contributed equally to this work.

✉ Thierry Etchegoyhen
tetchegoyhen@vicomtech.org

Harritxu Gete
hgete@vicomtech.org

¹ Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Donostia/San Sebastián, Gipuzkoa, Spain

² University of the Basque Country (UPV/EHU), Leioa, Bizkaia, Spain

Keywords Comparable corpora · Basque · Spanish · Neural Machine Translation

1 Introduction

Comparable corpora have long been identified as an important source of potentially useful data in natural language processing, notably to train data-driven machine translation systems (Munteanu & Marcu, 2005; Sharoff et al., 2016) or to extract bilingual dictionaries (Rapp, 1995).

Extracting sentence pairs with a significant degree of parallelism from this type of corpora faces important challenges, however, considering the vast amounts of unrelated or loosely related data that compose comparable datasets in different languages. Several methods have been designed to address the challenges of comparable document alignment (Azpeitia & Etchegoyhen, 2019; Sharoff et al., 2015) and comparable sentence alignment (Artetxe & Schwenk, 2019; Etchegoyhen & Azpeitia, 2016b; Fung & Cheung, 2004; Munteanu & Marcu, 2002; Smith et al., 2010), leading to new parallel datasets that can support machine translation, in particular for under-resourced languages (Etchegoyhen et al., 2016; Schwenk et al., 2019).

Neural Machine Translation (NMT) is currently the dominant paradigm in research and development in the field of Machine Translation, having led to significant advances in recent years for most language pairs (Bahdanau et al., 2015; Bojar et al., 2016, 2017; Vaswani et al., 2017). As a data-driven approach where model parameters are estimated and optimised on large volumes of parallel data, NMT is particularly sensitive to the presence of noise in the training data (Belinkov & Bisk, 2018; Cheng et al., 2018; Khayrallah & Koehn, 2018; Sperber et al., 2017). Due to the nature of the task, parallel data extracted from comparable corpora are likely to introduce unwarranted noise in the training process and an evaluation of this potential issue is worth examining.

In this article, we describe the results of an in-depth evaluation of the exploitation of a large strongly comparable corpus for Basque-Spanish, composed of independently-produced news by the Basque public broadcaster EITB,¹ focusing on the impact of various techniques to exploit the original data for Neural Machine Translation and extending the study by Etchegoyhen & Gete (2020).

We first compare results obtained with two alignment methods that have demonstrated high accuracy on comparable data, namely LASER (Artetxe & Schwenk, 2019) and STACC (Etchegoyhen & Azpeitia, 2016b), and evaluate various datasets and models based on the alignments obtained with each method via threshold filtering and data combination. We then measure the impact of filtering based on length difference statistical outliers between aligned source and target sentences, as comparable corpora typically feature information divergences between source and target sentences.

We also explore the impact of tags to identify comparable data in the training datasets, following the approach proposed by Caswell et al. (2019) for

¹ Euskal Irrati Telebista: <https://www.eitb.eus>.

backtranslations and Etchegoyhen & Gete (2020) for comparable data, and the use of mixed tagging by limiting tagging to portions of the data identified as containing more imbalanced information in the source and the target.

Finally, we compare the impact of using aligned comparable sentences to the use of backtranslations for NMT (Sennrich et al., 2016b), as both approaches have demonstrated their usefulness to increase the coverage and quality of translation models. We evaluate in particular the impact of mixtures of aligned and backtranslated data on the quality of the resulting translation models.

The remainder of this paper is organised as follows: Sect. 2 describes related work on comparable corpora; Sect. 3 describes the corpora, alignment methods and tools used in this study; in Sect. 4, we describe the baseline translation models based on the initial datasets; Sect. 5 centres on threshold filtering using the selected alignment methods; in Sect. 6, a detailed analysis of the different datasets is presented, while Sect. 7 describes the results obtained with different combinations of the available data; Sect. 8 describes the results obtained via length filtering and Sect. 9 those obtained via data tagging; in Sect. 10, we report on different experiments using backtranslated data; Sect. 11 presents a comparative summary of results on translation models; finally, Sect. 12 draws conclusion from the experiments and analyses.

2 Related work

A significant body of work has been produced over the years to mine and exploit parallel sentences from large collections of comparable monolingual corpora (Abdul-Rauf & Schwenk, 2009, 2011; Irvine & Callison-Burch, 2013; Munteanu & Marcu, 2005; Sharoff et al., 2016; Smith et al., 2010), starting with seminal work by Resnik (1999) to exploit the World Wide Web as a source of potential parallel data.

A standard part in the process is the determination of document pairs, to reduce the space of computation in the typically large datasets involved in the task. Several approaches have exploited metadata information for Web page alignment, including URL, structural tags or publication date (Chen & Nie, 2000; Munteanu & Marcu, 2005; Papavassiliou et al., 2016; Resnik & Smith, 2003). Alternatively, content-based document alignment approaches for comparable corpora have also been proposed, based on vector space models (Chen et al., 2004), token translation ratios (Ma & Liberman, 1999), mutual information (Fung & Cheung, 2004), expectation-maximisation (Ion et al., 2011) or n-gram matching using machine-translated documents (Uszkoreit 2010).

Several approaches have used a mixture of content and structural properties, notably the systems in the WMT 2016 document alignment shared task (Buck & Koehn, 2016a). Among those, Gomes and Lopes (2016) proposed a phrase-based approach combined with URL-matching, Buck and Koehn (2016b) used cosine similarity between TF/IDF vectors over machine-translated documents, and Germann (2016) performed alignment via vector space word representations, latent semantic indexing, and URL matching. In Esplá-Gomis et al. (2016), document alignment is

performed via a mixture of URL similarity, structural features such as shared links, and bag of words similarity.

Comparability at the document alignment was notably evaluated in a dedicated shared task of the BUCC workshop series (Sharoff et al., 2015). Among participating systems, Li and Gaussier (2013) used bilingual dictionaries and proportion of matching words to assess comparability, Morin et al. (2015) made use of hapax legomena and pigeon hole reasoning to enforce alignments, and Zafarian et al. (2015) used several components, including topic modelling, named entity detection and word features. A strictly content-based method was proposed by Etchegoyhen and Azpeitia (2016a), based on Jaccard similarity (Jaccard, 1901) over sets of lexical translations, expanded with surface-based entities and common prefix matching, demonstrating high accuracy in a large number of scenarios, including comparable corpora (Azpeitia & Etchegoyhen, 2019).

The extraction of parallel sentences from comparable corpora has also been addressed via a large variety of approaches, based on suffix trees (Munteanu & Marcu, 2002), maximum likelihood (Zhao & Vogel, 2002), binary classification (Munteanu & Marcu, 2005), cosine similarity (Fung & Cheung 2004), reference metrics over statistical machine translations (Abdul-Rauf & Schwenk, 2009; Sarikaya et al., 2009) or rich features (Smith et al., 2010; Ștefănescu et al., 2012). In recent years deep learning approaches have been applied to the task as well, via bidirectional recurrent neural networks (Grégoire & Langlais, 2017) or cosine similarity over bilingual sentence embeddings (Schwenk, 2018).

The core document alignment method of Etchegoyhen and Azpeitia (2016a) was applied to comparable sentence alignment, improving significantly over more sophisticated feature-rich methods (Etchegoyhen & Azpeitia, 2016b). This method, referred to as STACC, which is also based on Jaccard similarity over expanded lexical translation sets, was further extended with lexical weighting (Azpeitia et al., 2017) and a named-entity penalty (Azpeitia et al., 2018), obtaining the best results across the board in the BUCC shared task on parallel sentence identification in comparable corpora (Zweigenbaum et al., 2017, 2018).

Improvement over these results were obtained with the margin-based approach of Artetxe & Schwenk (2019), which uses cosine similarity over multilingual sentence embeddings, extended with a filtering mechanism based on nearest neighbours similarity. Large amounts of comparable corpora have been mined with this approach (Schwenk et al., 2019), by means of the LASER toolkit, which we include in our experiments on the selected corpora.

Recently, unsupervised approaches to NMT have demonstrated the potential of directly exploiting monolingual data to train neural translation models (Artetxe et al., 2018; Lample et al., 2018). A detailed comparison between this type of approach and the use of parallel sentences mined from comparable corpora may shed light on their respective merits, in isolation or in combination. Although this type of evaluation is outside the scope of the present study, the results in Sect. 10 on the use of back-translated data, compared to mining parallel sentences, provides preliminary results on the direct exploitation of monolingual data in this type of corpus.

Table 1 Original EITB corpus 2009–2018 in number of sentences (Sent.) and tokens (Tok)

Year	EU News	ES News	EU Sent.	ES Sent.	EU Tok.	ES Tok.
2009	18,552	18,759	236,753	223,323	3,745,794	5,378,433
2010	17,762	17,979	216,043	204,004	3,400,129	4,984,597
2011	18,856	19,037	230,902	216,240	3,771,173	5,698,541
2012	19,272	18,903	229,270	213,730	3,761,079	5,706,086
2013	13,520	13,449	180,552	174,752	2,966,897	4,585,870
2014	16,249	15,684	217,314	190,722	3,672,011	5,356,644
2015	13,906	13,433	196,161	170,929	3,279,763	4,724,633
2016	16,288	16,654	211,899	192,563	3,454,514	5,100,300
2017	16,946	19,178	218,123	211,131	3,534,538	5,478,081
2018	17,633	21,272	235,419	246,261	3,839,636	6,196,128
Total	168,984	174,348	2,172,436	2,043,655	35,425,534	53,209,313

A first version of the EITB corpus was prepared and shared with community (Etchegoyhen et al., 2016), to provide further support to the under-resourced Basque-Spanish language pair. In Etchegoyhen & Gete (2020), a new version of the corpus was prepared and shared with the scientific community,² built with news generated in subsequent years. In what follows, we use the latter version of the original data, covering news produced between 2009 and 2018, and evaluate the impact of different methods to exploit these comparable data on Neural Machine Translation.

3 Corpora preparation

The EITB corpus is composed of news independently produced in Basque and Spanish by the journalists of the Basque Country's public broadcast service, to report on the same specific events. The corpus can thus be considered strongly comparable (Skadina et al., 2012) and viewed as a rich source of parallel data for this language pair (Etchegoyhen et al., 2016).

This version of the original dataset covers 10 years of content, and is composed of 168,984 documents in Basque and 174,348 in Spanish, extracted from XML files.³ With an original amount of over two million sentences per language, it is one of the largest available corpora for the Basque-Spanish language pair, covering political news, sports and cultural events, among others. Statistics of the original corpus are described in Table 1.

To perform the experiments described in this article, a second corpus was used, based on the translation memories made available by the Basque government as

² The corpus is available in the OPUS repository (Tiedemann, 2012): <https://opus.nlpl.eu/EiTB-ParCC.php>

³ See Etchegoyhen et al. (2016) for a detailed description of the original file structure.

Table 2 Corpora statistics (number of sentence pairs)

CORPUS	ODE	EITB-STACC 0,0	EITB-LASER 0,0
TRAIN	921,763	1,337,040	1,383,309
DEV	2000	2000	2000
TEST	2000	1678	1678

open data.⁴ We used the corpus created from these translation memories and shared by Etchegoyhen et al. (2018),⁵ to which we will refer as ODE in what follows. The corpus is constituted mainly of translations of administrative content, notably from the Instituto Vasco de Administración Pública (IVAP). As it consists of professionally produced translations, the ODE corpus was used as a parallel basis for the experiments described in the remainder of this work.

The EITB corpus was first aligned at the document level with DOCAL, the efficient component for parallel and comparable document alignment described in Etchegoyhen and Azpeitia (2016a) and Azpeitia and Etchegoyhen (2019). For sentence alignment, we used the two previously mentioned approaches:

- STACC (Etchegoyhen & Azpeitia, 2016b): This method is based on sets of bilingual lexical mappings, set expansion operations to include unseen named entities determined by surface form as well as longest common prefixes, and Jaccard similarity (Jaccard, 1901) over the source to target and target to source set. We use the STACC variant with lexical weighting (Azpeitia et al., 2017), with lexical weights computed on the corpus to be aligned.
- LASER (Artetxe & Schwenk, 2019): This approach is based on multilingual sentence embeddings, cosine similarity over embedding pairs, and a margin computed over closest alignment candidates.

To extract the translation tables necessary for both the DOCAL and STACC approaches to document and sentence alignment, we used the FASTALIGN toolkit (Dyer et al., 2013) on the ODE corpus. Prior to performing either alignment, the sentences were tokenised and truecased using the scripts available in the MOSES toolkit (Koehn et al., 2007).

The statistics for the ODE and EITB corpora, in number of aligned sentences are shown in Table 2. For the EITB corpus, alignments results are based on one-to-one alignments without any filtering based on the alignment scores computed by STACC or LASER, i.e. with an alignment threshold set to 0 in both cases.

The development and test sets are random partitions for both corpora; for the EITB corpus, the extracted alignments were manually verified. For the test set, additional Basque translations were manually created by professional translators, to increase the robustness of the test against the relatively free word order in Basque (Etchegoyhen et al., 2018). In the Spanish to Basque translation direction, we thus

⁴ <https://opendata.euskadi.eus>.

⁵ In the version shared for the IWSLT 2018 shared task, available at: <https://sites.google.com/site/iwslt-evaluation2018/TED-tasks>.

use two translation references, whereas for Basque to Spanish, we use the manually translated sentences as source, to remove any eventual bias that may arise from the original data being comparable.

4 Baseline models

All translation models were based on the Transformer architecture (Vaswani et al., 2017), built with the MARIAN toolkit (Junczys-Dowmunt et al., 2018). The models consisted of 6-layer encoders and decoders, feed-forward networks of 2048 units, embeddings vectors of dimension 512 and 8 attention heads. The dropout rate between layers is 0.1.

We used the Adam optimiser with $\alpha = 0.0003$, $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$. The learning rate increases linearly for the first 16,000 training steps and decreases thereafter proportionally to the inverse square root of the corresponding step. We set the working memory to 6000MB and automatically chose the largest mini-batch that fit the specified memory. The validation data was evaluated every 3500 steps, and the training process ended if there was no improvement in the perplexity of 10 consecutive checkpoints. Embeddings for source, target and output layer are tied and all datasets were segmented with BPE (Sennrich et al., 2016c), using 30,000 operations.

We trained five baseline models, first on the ODE corpus and on that same corpus merged with the aligned EITB corpus, with an alignment threshold set to 0 for both the STACC and LASER methods, to measure the initial contribution of the selected comparable data. Additionally, we trained models only on the aligned EITB datasets, to measure the impact of the comparable data in isolation. These models and all subsequent ones were evaluated on the previously described test sets, which cover various topics. The results in terms of BLEU (Papineni et al., 2002) for the baseline models are shown in Table 3 and were computed with SacreBLEU (Post, 2018).⁶

In both translation directions, the contribution of the comparable data was significant, with large improvements in BLEU scores on the in-domain test sets. Although this is not unexpected, since the ODE model was trained on data in the administrative domain, this confirms the potential of the strongly comparable EITB data as a means to improve Basque-Spanish translation, as previously established in Etchegoyhen et al. (2016). Taken in isolation, the comparable data provided strong baseline results on the EITB test sets, but perform poorly on the out-of-domain ODE test sets. Overall, the combination of EITB and ODE data was optimal in both translation directions, when considering average results on the in-domain and out-of-domain test sets.

The initial alignments obtained with LASER and STACC resulted in similar improvements across the board, except for Spanish to Basque translation where the former performed better on the in-domain EITB test set. In the next sections, we evaluate several variants of the initial alignments, both individually and in

⁶ For all translation model results hereafter, translation from Basque to Spanish is indicated as EU-ES and the reverse direction as ES-EU.

Table 3 BLEU results for the baseline models in both translation directions

Model	ES-EU		EU-ES	
	Test EITB	Test ODE	Test EITB	Test ODE
ODE	28.6	32.4	34.0	46.7
EITB-STACC _{0,0}	34.7	11.3	44.8	24.5
EITB-LASER _{0,0}	36.4	12.0	44.9	25.1
ODE + EITB-STACC _{0,0}	36.1	30.6	46.5	44.7
ODE + EITB-LASER _{0,0}	36.9	30.2	46.8	45.0

combination, to determine an optimal setup for the exploitation of the original comparable dataset.

5 Threshold filtering

Methods to extract parallel data from comparable corpora rely on metrics that measure translation equivalence in some form. The scores assigned by the core metrics usually need to be complemented with some form of filtering, based on thresholds determined over the training or development datasets (Artetxe & Schwenk, 2019; Etchegoyhen & Azpeitia, 2016b). This is necessary since similarity varies between comparable sentences and comparability metrics usually assign a continuous score to comparable sentence pairs.

To determine the impact of threshold selection, we extracted subsets of the aligned EITB corpus after applying different thresholds, selected according to the ranges of the STACC and LASER metrics, which produced significant amounts of initial data. Each subset was then used in combination with the ODE corpus, to train new NMT model variants.

Tables 4 and 5 show the results, in terms of BLEU score over the EITB test sets and number of aligned sentences, for the threshold-filtered datasets with STACC and LASER, respectively. For each method, we computed statistical significance via bootstrap resampling (KoeHN, 2004) between each filtered variant and the 0.0 baseline.⁷

In the case of STACC, neither the slight increases over the baseline in ES-EU nor the slight decreases in EU-ES up to threshold 0.17 were statistically significant. For further experiments, either dataset, up to a threshold of 0.20 in the first case and 0.17 in the latter case, could thus be considered equivalent, at least regarding the training of the complete models that include the ODE datasets as well.

Higher STACC thresholds usually increase the accuracy of the aligner when measured in terms of precision and recall on various alignment tasks (Azpeitia et al., 2017; Etchegoyhen & Azpeitia, 2016b; Etchegoyhen et al., 2016), or when training NMT models via fine-tuning over aligned datasets (Etchegoyhen & Gete, 2020), and could have been selected accordingly. Our goal in these experiments,

⁷ In both tables, † indicates statistical significance over the baseline, for $p < 0.05$.

Table 4 Alignment threshold results with STACC

Threshold	Aligned	BLEU	
		ES-EU	EU-ES
0.0	1,337,040	36.1	46.5
0.15	1,122,890	36.4	46.4
0.17	930,839	36.4	46.4
0.20	580,478	36.7	45.7 [†]
0.25	181,405	33.4 [†]	41.4 [†]

Table 5 Alignment threshold results with LASER

Threshold	Aligned	BLEU	
		ES-EU	EU-ES
0.0	1,383,309	36.9	46.8
1.0	1,096,190	37.8 [†]	46.9
1.03	740,544	38.3 [†]	46.5
1.04	638,633	38.5 [†]	46.4
1.05	541,900	38.5 [†]	45.9 [†]
1.06	450,176	38.3 [†]	45.7 [†]

however, was to measure the impact of complementing parallel data with comparable data, in a realistic scenario where different datasets are merged to maximise NMT models model coverage and accuracy. We therefore opted for an initial data selection based on the statistical significance of BLEU scores for NMT models trained on different subsets. To measure the impact of additional filtering methods, we thus selected the largest of the equivalent datasets in terms of impact on BLEU scores, i.e. the 0.0 baseline.

With LASER, all thresholds led to statistically significant gains over the baseline in ES-EU, whereas none did for EU-ES. Additionally, we computed statistical significance for all improving variants against the model based on the 1.0 threshold, with no significance for $p < 0.05$. Along the same lines as those that guided the selection of the STACC dataset, for the next experiments we selected the datasets that led to the best LASER-based models, while also featuring the largest amount of data. Thus, for ES-EU we selected as defaults the dataset based on the 1.0 threshold, and for EU-ES the dataset based on the 0.0 threshold.

Table 6 summarises the main characteristics of the selected datasets for each method and language pair.

6 Data analysis

Each alignment method generates its own datasets according to the selected threshold, which may differ in terms of several characteristics, while also overlapping to a certain extent. In this section, we analyse the main differences in terms of corpus statistics.

Table 6 Selected datasets for ES-EU and EU-ES based on threshold results

LANG	METHOD	THRESHOLD	ALIGNED
ES-EU	STACC	0.0	1,337,040
ES-EU	LASER	1.0	1,096,190
EU-ES	STACC	0.0	1,337,040
EU-ES	LASER	0.0	1,383,309

Table 7 Number of aligned sentences for the union and intersection of the STACC and LASER datasets

LANG	DATASET	ALIGNED
ES-EU	STACC _{0.0} $\cup$ LASER _{1.0}	1,563,998
	STACC _{0.0} $\cap$ LASER _{1.0}	869,234
EU-ES	STACC _{0.0} $\cup$ LASER _{0.0}	1,809,659
	STACC _{0.0} $\cap$ LASER _{0.0}	909,106

We first extracted the union and intersection of the STACC and LASER datasets, with results shown in Table 7. Overall, the selected methods mined a common core of 55.6% of aligned sentences for ES-EU, and 50.2% for EU-ES. This leaves an almost equivalently large amount of alignments uniquely mined by each method.

Determining the benefits these unique alignments can provide will be approximated in the next section by training models based on the union and intersection of the datasets. A full manual analysis of alignment accuracy for these large datasets was not feasible for this study, therefore we randomly sampled 100 aligned sentences from each one of the selected datasets and manually evaluated the alignments for approximate estimates of the quality of the unique alignments. An alignment was considered incorrect if there was information imbalance between source and target sentences, or crucial information missing, such as Named Entities.

For EU-ES, where both LASER and STACC datasets are based on 0 thresholds, the number of alignments identified as correct in the samples was 21% for the former, and 35% for the latter. For ES-EU, the STACC dataset is the same as in the other direction, but the LASER dataset is based on a 1.0 threshold for which we extracted a separate sample. In this case, 29% of the alignments were considered correct.

Although these results are only approximate estimates, they do indicate that each method addresses aspects of comparable data alignment that the other method fails to capture, while also featuring erroneous alignments based on the respective characteristics of the methods. From manual examination of the samples, the main weakness of LASER seems to be the lack of robustness in identifying corresponding Named Entities and alignment of loosely related sentences, which is likely to be a by-product of the use of embeddings, whereas for STACC the main limitation appeared to be the alignment of sentences which are translations of each other, but where a significant portion of one side of the pair is missing in the other.

We provide examples of correct alignments in Table 8. As these examples illustrate, each method can mine valid parallel sentences in the comparable datasets that are not captured by the alternative, irrespective of the length of the aligned sentences.

Table 8 Examples of sentence pairs uniquely aligned by either STACC or LASER, with corresponding English translations

METHOD	LANG	EXAMPLE
STACC	ES	el ambiente será algo más frío
	EU	giroa zertxobait hoztuko da
	EN	<i>the atmosphere will be a bit colder</i>
	ES	el presidente del Araba Buru Batzar responderá a las preguntas de Aitzol Zubizarreta y comentará diversas cuestiones de la actualidad política vasca
	EU	Araba Buru Batzarreko presidentek Aitzol Zubizarretaren galderari erantzungo die gaurkotasun politikoa osotzen duten hainbat gai komentatuz
LASER	EN	<i>Araba Buru Batzar president will answer questions from Aitzol Zubizarreta and comment on various matters of current Basque politics</i>
	ES	es una noticia falsa
	EU	albiste faltsua da
	EN	<i>it's fake news</i>
	ES	según los paleontólogos sudafricanos, “ estos indicios muestran que los homínidos conocían la tecnología del fuego hace cerca de un millón de años ”
	EU	Hegoafrikako paleontologoek ustez , “ hominidoek jada suaren teknologia orain milioi bat urte inguru ezagutzen zutela adierazten du aurkikuntza honek ”
	EN	<i>according to South-African paleontologists, “these clues show that hominids knew about fire technology about a million years ago”</i>

Fig. 1 Vocabulary distribution (number of tokens) over selected ES-EU datasets

Fig. 2 Vocabulary distribution (number of tokens) over selected EU-ES datasets

We then compared the vocabulary distribution differences between the selected datasets, the union of the two, and their intersection. Figures 1 and 2 show the number of tokens in the union vocabulary sizes for ES-EU and EU-ES, respectively, along with the amount of common tokens between the union and intersection datasets, the STACC dataset and the LASER dataset.

Table 9 Number (#) and percentage (%) of aligned named entities (NE) and identical sentences (ID)

LANG	DATASET	NE		ID	
		#	%	#	%
ES-EU	STACC _{0,0} $\cup$ LASER _{1,0}	243,888	–	69,483	–
	STACC _{0,0}	234,374	96.10	65,457	94.2
	LASER _{1,0}	218,342	89.53	68,893	99.2
EU-ES	STACC _{0,0} $\cup$ LASER _{0,0}	261,548	–	69,483	–
	STACC _{0,0}	234,374	89.61	65,457	94.2
	LASER _{0,0}	250,675	95.84	68,893	99.2

The intersection of vocabularies is lesser than the union, indicating that each method mines portions of data that differ in terms of vocabulary, with 85% common vocabulary for both ES-EU and ES-EU .

As indicated by the STACC-only and LASER-only results, uniquely retrieved tokens range between 40,707 and 87,504 for EU-ES, with larger counts for LASER, and between 33,457 and 83,577 for ES-EU, with larger counts for STACC in this case. Considering the reversed tendencies given the language pair, these differences are likely due in part to the differences in volumes of selected data, with 46,209 more sentence pairs selected by LASER for EU-ES, and 240,731 more by STACC for ES-EU. The differences in vocabulary indicate however that the additional sentence pairs in each case differ significantly from the common pairs selected by both methods.

Since the two methods differ significantly in the way they mine parallel sentences, via bilingual embeddings in one case and lexical translation overlap in the other, the differences in terms of vocabulary are not surprising. To further explore these differences, we measured the amount of Named Entities (NE) and sentence pairs that are identical in each language, as retrieved by each method.⁸ Results are shown in Table 9.

In terms of NEs, the results are in line with the previously discussed volume differences between the selected datasets, with either methods retrieving at least 9 out of 10 NEs present in the union of the datasets.

The methods do contrast in terms of identical sentences in aligned pairs, with LASER aligning virtually all of them in both cases, as opposed to STACC. This difference might be due to cases where the identical source and target sentences are in the same language, in which case the reliance of the STACC method on lexical translation mappings would discard these pairs as possible alignments, correctly so in these cases. Since the amount of such pairs is comparably marginal, we did not explore these cases any further.

Alignment methods may also differ in terms of the length of the sentences in aligned pairs. We measured both the average sentence length and the average length difference between source and target sentences, with the results shown in Table 10.

⁸ The identification of Named Entities was approximated by determining the amount of cased tokens and numbers in the truecased datasets. Identical sentences in the source and target may be due to alignment errors but also to sports results or movie listings, for example.

Table 10 Average sentence length and sentence length differences (number of tokens) for pairs aligned with STACC or LASER

	Average sentence length		Average length difference
	ES	EU	ES-EU
STACC 0.0	29.28	18.39	11.64
LASER 0.0	27.41	18.50	9.44
LASER 1.0	27.33	18.79	9.00

As indicated by these results, STACC tends to align pairs with larger length differences than LASER, with a two-token difference on average. This is mainly due to the length of the selected Spanish sentences, as the length of Basque sentences is comparable with both methods on average. These differences add to the observation that the characteristics of each method lead to divergent mining of parallel sentences in comparable corpora, to some extent. In the next section, we investigate the impact of using different combinations of data to train NMT models.

7 Data combination

To measure the impact of combining the datasets generated by both STACC and LASER, we created variants of the corpus with both the union and the intersections of these datasets and trained complete models with each variant.

Table 11 presents the results in terms of BLEU scores for ES-EU, along with the previous scores obtained with the individual datasets selected by each method. Table 12 shows statistical significance results of the scores obtained between each pair of models.

The union of the two datasets results in the lowest score overall for ES-EU, which is statistically significant against all other models, with differences in BLEU scores ranging from 1 to 2.7 points. Although each method can uniquely retrieve valid alignments on its own, as previously discussed, in combination the results were lower on these test sets. Although this may indicate a tendency, additional test sets might be warranted to further confirm if the union of alignment pairs is always detrimental for this translation pair. It should be noted in particular that translation into Basque is more likely to be sensitive to variations in word order between valid translations and fixed references, variations which are likely to increase with larger amounts of training data.

The intersection-based model was not significantly different from the LASER-only variant, although it features lower amounts of data. The gains obtained over the STACC model were significant, indicating that the intersection-induced filtering was beneficial for part of the data selected by STACC.

Tables 13 and 14 present the results for EU-ES, in terms of BLEU and statistical significance, respectively.

For this translation pair, the model trained on the union of the datasets produced results on a par with those trained on the datasets mined by either STACC or LASER,

Table 11 BLEU scores in ES-EU and number of aligned EITB sentences with union and intersection datasets

DATASET	ALIGNED	BLEU
ODE + STACC _{0.0}	1,337,040	36.1
ODE + LASER _{1.0}	1,096,190	37.8
ODE + STACC _{0.0} $\cup$ LASER _{1.0}	1,563,998	35.1
ODE + STACC _{0.0} $\cap$ LASER _{1.0}	869,234	37.6

Table 12 Statistical significance results in ES-EU with bootstrap resampling for models trained on ODE+STACC0.0 and ODE+LASER1.0 (indicated in the table as STACC and LASER, respectively)

	STACC	LASER	STACC $\cup$ LASER	STACC $\cap$ LASER
STACC	—	0	0.002	0.006
LASER	0	—	0	0.102
STACC $\cup$ LASER	0.002	0	—	0
STACC $\cap$ LASER	0.006	0.102	0	—

Table 13 BLEU scores in EU-ES and number of aligned EITB sentences with union and intersection datasets

DATASET	ALIGNED	BLEU
ODE + STACC _{0.0}	1,337,040	46.5
ODE + LASER _{0.0}	1,383,309	46.8
ODE + STACC _{0.0} $\cup$ LASER _{0.0}	1,809,659	46.2
ODE + STACC _{0.0} $\cap$ LASER _{0.0}	909,106	42.1 [†]

Table 14 Statistical significance results in EU-ES with bootstrap resampling for models trained on ODE+STACC0.0 and ODE+LASER0.0 (indicated in the table as STACC and LASER, respectively)

	STACC	LASER	STACC $\cup$ LASER	STACC $\cap$ LASER
STACC	—	0.183	0.170	0
LASER	0.183	—	0.081	0
STACC $\cup$ LASER	0.170	0.081	—	0
STACC $\cap$ LASER	0	0	0	—

with no statistically significant difference. In contrast, the intersection-based model was significantly worse than all other variants, indicating that lower amounts of training data were more impactful for this translation pair than was the case for ES-EU. In the following sections, we further explore the differences between the two translation pairs.

8 Length filtering

Due to the nature of the task, aligned comparable sentences may display information imbalance, with one of the sentences in a pair missing part of the information in the other. In this section we evaluate the impact of information mismatch, via filtering based on length differences measured on the aligned sentence pairs.

We based our approach to length-based filtering on the method described in Etchegoyhen et al. (2018), which aims to identify statistical outliers in terms of length differences between aligned sentences. We first computed the median and standard deviation over length differences, measured in terms of tokens. These reference statistics were computed on the parallel ODE corpus, to establish the relevant length-difference indicators on parallel human translations. A length-difference score (LGS), based on a modified z-score, was then computed on the aligned EITB datasets, according to the formula in Eq. 1:

$$LGS = \frac{0.6745 \times (x - \bar{y})}{median(\{|y_i - \bar{y}|\})}, \quad (1)$$

where x is the length difference of a sentence pair in the EITB corpus, $\bar{y}$ is the median length difference in the reference corpus, and the denominator is the median absolute deviation, computed over the reference corpus as well.

The modified z-score was then used to identify outliers in the aligned EITB corpus, with sentence pairs having an absolute score over a given threshold identified as cases of information imbalance. Iglewicz and Hoaglin (1993) recommend a value of 3.5 to identify outliers when using a modified z-score, and we selected this value as our default to filter all identified outliers. Additionally, we selected two more thresholds with lower values, namely 2.0 and 1.5, to evaluate the impact of a more restrictive identification of length imbalance.⁹

Table 15 summarises the amount of sentence pairs discarded with the above length filtering method, with different thresholds. Overall, the volumes of discarded data are significant, even with less restrictive thresholds, with a minimum of 25% of the data for the selected datasets with LGS_{2.0} filtering, for instance. Higher alignment threshold gradually minimise the issue in both approaches, although it tends to persist throughout.

These results also reinforce one of the conclusions of the manual analysis of samples, namely that the STACC method tends to align more sentences than LASER where part of the information is a correct translation, but a significant part of it is missing on one side of the pair.

In Table 16, we show examples of filtered pairs, where the information in the Spanish sentence that is missing in the Basque counterpart of the aligned pair is marked in italics.

⁹ Length imbalance could have been computed by simply taking the average absolute difference for each sentence pair. However, this method would not lead to the identification of statistically significant deviations from the mean determined on a reference corpus, which was our goal for these experiments.

Table 15 Amount of sentence pairs discarded via length filtering with LASER and STACC

DATASET	ALIGNED	LGS _{3,5}	LGS _{2,0}	LGS _{1,5}
STACC 0.0	1,337,040	283,392 (21.20%)	485,478 (36.30%)	565,613 (42.3%)
STACC 0.15	1,122,890	219,294 (19.53%)	388,672 (34.61%)	457,245 (40.72%)
STACC 0.17	930,839	158,822 (17.06%)	297,339 (31.94%)	355,015 (38.14%)
STACC 0.20	580,478	64,566 (11.12%)	142,827 (24.61%)	177,818 (30.63%)
STACC 0.25	181,405	4646 (2.56%)	14,993 (8.26%)	20,881 (11.51%)
LASER 0.0	1,383,309	198,894 (14.38%)	388,918 (28.12%)	470,639 (34.02%)
LASER 1.0	1,096,190	144,141 (13.15%)	293,601 (26.78%)	358,662 (32.72%)
LASER 1.03	740,544	72,185 (9.75%)	171,594 (23.17%)	217,188 (29.33%)
LASER 1.04	638,633	53,454 (8.37%)	137,273 (21.49%)	176,772 (27.68%)
LASER 1.05	541,900	38,005 (7.01%)	106,321 (19.62%)	139,609 (25.76%)
LASER 1.06	450,176	25,522 (5.67%)	78,798 (17.50%)	105,847 (23.51%)

The results on models trained on selected corpora filtered by length outliers for ES-EU and EU-ES are shown in Tables 17 and 18, respectively.¹⁰ Also indicated in the tables are the size of the filtered corpus, the BLEU brevity penalty (BP), and the proportion of filtered sentences where the length of the Spanish sentence is larger than that of the Basque sentence. Statistical significance was computed for each model with respect to the unfiltered baselines.

For Spanish to Basque translation, in terms of BLEU scores, length filtering improved significantly over the unfiltered corpus for all variants of the union dataset and for the intersection dataset, where the results were statistically significant with the less restrictive of the three filtering thresholds. For this language pair, information imbalance was thus significantly detrimental in the datasets unfiltered for length differences. For Basque to Spanish, the results were reversed, with a gradual decrease of BLEU scores with additional filtering, all statistically significant.

One interpretation of these opposite results may be based on the fact that the length of filtered Spanish sentences is systematically longer than that of Basque sentences. Although this is the case in general, given the morphological system of Basque with productive affixation, more aggressive filtering of length-difference outliers lowers the proportion of Spanish sentences that are longer than their Basque counterparts, as shown in the last column of the tables, indicating that the overall tendency in the corpus is for information imbalance to affect the Basque data more than its Spanish counterpart. In other words, the news in the EITB corpus tend to summarise the information more in Basque than in Spanish. Translating from Spanish to Basque would thus have the effect of orienting the models towards summarisation, with a negative impact on translation quality that needs to be compensated with more length-based filtering. This conjecture is supported by the

¹⁰ For Basque to Spanish translation, we discarded the intersection dataset, considering the significantly worse results obtained over the baseline, as discussed in the previous section.

Table 16 Examples of filtered sentence pairs with modified z-score above 2.0 for STACC and LASER, with English translations

METHOD	LANG	EXAMPLE
STACC	ES	la web traducida a 6 idiomas , incluyendo nuevos mercados como el chino , ofrece la oportunidad de acceder a la información más novedosa sobre las empresas de la CAV en el mundo
	EU	webgune hori sei hizkuntzataraz itzultzen da , txinerara tartean Txinako merkatu berriak barne hartzearren
	EN	the web translated to 6 languages, including new markets like Chinese, presents the opportunity to access new information on CAV companies worldwide
	ES	si más de un incondicional consiguiera llegar al final del visionado , se pondrá en marcha la segunda fase del concurso , que consiste en un test de preguntas sobre la serie
LASER	EU	bat baino gehiago emanaldi amaterara iristen bada lehiaketaren bigarren epea jarriko da martxan
	EN	If more than one unqualified manages to reach the end of the viewing, the second phase of the contest will be launched, which consists of a question test about the series
	ES	Madariaga no ha precisado cuántos de los trabajadores de la planta TMB resultaron afectados por la fiebre Q , dado que es el Departamento de Salud del Gobierno Vasco el que debe hacer públicos aún los resultados , y ha reiterado el dato ya conocido de que hay ocho casos confirmados y que hubo 34 empleados de baja , aunque se sabe por la empresa que la cifra “ ha ido en aumento ”
	EU	Q sukarrak TMB lantegiko zenbat langileri eragin die ez du zehaztu Madariagak ; izan ere , Eusko Jaurlaritzako Osasun Sailak emaitzak agitaratu behar ditu oraindik
	EN	Madariaga has not specified how many of the workers at the TMB plant were affected by Q fever, since it is the Health Department of the Basque Government that has yet to make the results public, and has reiterated the already known data that there are eight cases confirmed and that there were 34 employees on leave, although it is known by the company that the number “has been increasing.”
	ES	es a la presidenta del Congreso a quien corresponde fijar el día y la hora de comienzo del debate de investidura , con la vista puesta en el límite del lunes 31 de octubre , cuando vence el plazo para que se haya investido un presidente del Gobierno ; de lo contrario , el rey tendría que convocar terceras elecciones para diciembre
	EU	Kongresuko presidenteari dagokio inbestidura-saioaren data eta ordua zehaztea , urriaren 31 izanik epemuga
	EN	It is up to the President of Congress to set the day and time for the inauguration debate to begin, with an eye on the limit of Monday, October 31, when the deadline for a Prime Minister to be invested has expired; otherwise, the king would have to call third elections for December

Table 17 Results from length filtering for ES-EU on union and intersection datasets

	ALIGNED	BLEU	BP	$\text{len}(\text{filt}_{es}) > \text{len}(\text{filt}_{eu})$ (%)
ODE + STACC $_{0.0} \cup$ LASER $_{1.0}$	1,563,998	35.1	0.801	–
ODE + STACC $_{0.0} \cup$ LASER $_{1.0} + \text{lgs}_{3.5}$	1,259,125	39.4 [†]	0.938	98.96
ODE + STACC $_{0.0} \cup$ LASER $_{1.0} + \text{lgs}_{2.0}$	1,036,906	39.3 [†]	0.974	98.42
ODE + STACC $_{0.0} \cup$ LASER $_{1.0} + \text{lgs}_{1.5}$	947,302	38.6 [†]	0.985	98.11
ODE + STACC $_{0.0} \cap$ LASER $_{1.0}$	869,234	37.6	0.877	–
ODE + STACC $_{0.0} \cap$ LASER $_{1.0} + \text{lgs}_{3.5}$	749,667	39.1 [†]	0.945	99.60
ODE + STACC $_{0.0} \cap$ LASER $_{1.0} + \text{lgs}_{2.0}$	622,983	38.4	0.975	99.37
ODE + STACC $_{0.0} \cap$ LASER $_{1.0} + \text{lgs}_{1.5}$	568,575	38.6	0.981	99.24

Table 18 Results from length filtering for EU-ES on the union dataset

	ALIGNED	BLEU	BP	$\text{len}(\text{filt}_{es}) > \text{len}(\text{filt}_{eu})$ (%)
ODE + STACC $_{0.0} \cup$ LASER $_{0.0}$	1,809,659	46.2	0.934	–
ODE + STACC $_{0.0} \cup$ LASER $_{0.0} + \text{lgs}_{3.5}$	1,465,775	45.5 [†]	0.930	99.01
ODE + STACC $_{0.0} \cup$ LASER $_{0.0} + \text{lgs}_{2.0}$	1,209,552	45.1 [†]	0.926	98.51
ODE + STACC $_{0.0} \cup$ LASER $_{0.0} + \text{lgs}_{1.5}$	1,105,465	44.2 [†]	0.929	98.21

results in terms of brevity penalty, with lower brevity scores correlating with less length-based filtering.

For Basque to Spanish, translation quality seems to correlate instead with the volumes of data. This may be attributed to the fact that there is no marked tendency towards summarisation in this translation direction, given the fact that the target sentences are longer than the source, for the most part. The target monolingual data can thus contribute relevant decoding information in a way that is similar to synthetic data based on back-translations or on empty source sentences (Sennrich et al., 2016b), where the models can improve their modelling of the target sequences in the face of degenerate source input.

9 Data tagging

The use of tags identifying specific aspects of the data in the training corpora has proved effective in Neural Machine Translation. Thus, Sennrich et al. (2016a) used markers to control the translation of honorifics, Kobus et al. (2017) model domain control via tags identifying different domains, Yamagishi et al. (2016) use tags to control voice translation in Japanese to English Translation, and Caswell et al. (2019) employ tags to identify back-translated synthetic data, for instance.

The latter work in particular demonstrates that tagging techniques can prove more effective than noising approaches, indicating also that the impact of noising

for back-translated data essentially acts as an indicator of the type of data used for training and helps the models discriminate between natural and synthetic data. We extend their approach to comparable data, by prepending a <cc> tag to all source sentences of the comparable EITB training set.

We trained models by combining the ODE corpus with selected variants of the EITB corpus, with and without tags indicating comparable data, on the best performing datasets as established in the previous sections. The results of these experiments are shown in Table 19.

For Spanish to Basque, tagging was only effective on the noisier dataset, i.e. the EITB variant with no filtering of length-difference outliers. For the less noisy dataset, based on length-based filtering, the use of tags was detrimental. Interestingly, the use of tags in this translation direction had a significant impact in terms of shortness of translations, from a brevity penalty of 0.801 for the untagged model to 0.997 for the tagged variant based on the same corpus.

These results tend to support the hypothesis that tagging helps the model discriminate between natural and noisy data, and becomes counterproductive when the tagged comparable data are closer to natural translations, as in the variant based on filtered length-imbalanced data.

For Basque to Spanish, tagging was detrimental in terms of BLEU, despite minor improvements regarding the brevity of translations. This can be viewed in light of the previous hypothesis that the overall higher quantity of information in the Spanish target sentences is a dominating factor for this translation direction.

The negative impact caused by tagging in this case seems to indicate that comparable data with less source information in the source than in the target are actually not noisy for the translation models, as discriminating between natural and comparable data leads to lower translation quality results in this case. Determining whether this hypothesis is correct can be further examined by comparing tagged back-translated data with tagged comparable data; we explore this hypothesis in the next section.

As a final experiment with tagged data, we evaluated a mixed use of tags for ES-EU, the language pair for which tagging was beneficial. Given that tags were effective for the noisier dataset without length filtering and not for the filtered variant, we tagged only the data discarded via length-filtering, leaving the non-filtered data without tagging. The results are shown in Table 20.

Mixed tagging obtained degraded results over generalised tagging, indicating that, for this language pair, identifying all comparable data in the training datasets remains the optimal option. A more fine-grained analysis might be needed to further explore the use of mixed tagging; we leave this type of evaluation for future research.

10 Backtranslations vs. alignment

Backtranslations have proved useful to complement parallel corpora with synthetic translations (Edunov et al., 2018; Poncelas et al., 2018; Sennrich et al., 2016b). To our knowledge, no comparison has yet been made between using the target side of

Table 19 BLEU scores on merged datasets with and without tags

LANG	MODEL	NO-TAG		TAG	
		BLEU	BP	BLEU	BP
ES-EU	ODE + STACC _{0.0} $\cup$ LASER _{1.0}	35.1	0.801	39.7 [†]	0.997
	ODE + STACC _{0.0} $\cup$ LASER _{1.0} + lgs _{3.5}	39.4	0.938	39.2 [†]	0.992
EU-ES	ODE + STACC _{0.0} $\cup$ LASER _{0.0}	46.2	0.934	45.6 [†]	0.971

Table 20 BLEU scores for ES-EU with different tagging methods on the ODE + STACC_{0.0} $\cup$ LASER_{1.0} dataset

	BLEU	BP
NO-TAG	35.1	0.801
TAG	39.7 [†]	0.997
MIXED-TAGGING	38.4 [†]	0.927

Table 21 Comparative results on the EITB test sets using backtranslations merged with ODE data

MODEL	ES-EU		EU-ES	
	BLEU	BP	BLEU	BP
ODE	28.6	1.000	34.0	0.980
BEST MERGED	39.7 [†]	0.997	46.2 [†]	0.934
ODE + BT	35.8 ^{†‡}	1.000	44.8 ^{†‡}	1.000
ODE + BT TAGGED	36.3 ^{†‡}	1.000	44.2 ^{†‡}	0.991

The best variant from the merged corpora is indicated as BEST MERGED. Models that include backtranslations are indicated with BT. Statistical significance ($p < 0.05$) over the ODE baseline is indicated with [†] and significance over the BEST MERGED model with [‡].

comparable corpora via backtranslations vs. using the results of alignment, as is typically done to exploit this type of corpora and as we have been adopting so far in this work. Given the results obtained in the previous sections, for translation from Basque to Spanish in particular, it might be the case that, in some cases, backtranslating target data provides similar or larger improvements over the baselines than aligned data.

To explore this hypothesis, we generated backtranslations for both translation directions, using the baseline ODE model, and trained translation models by merging the resulting backtranslations with the ODE corpus. We also trained a model where backtranslated data were marked with a <bt> tag Caswell et al. (2019). The results are shown in Table 21.

Table 22 Comparative results with mixed aligned and backtranslated data

	MODEL	BLEU	BP
	ES-EU	BEST MERGED	39.4
		BEST MERGED + BT	38.9
		BEST MERGED TAGGED	39.7
		BEST MERGED TAGGED + BT	39.6
Statistical significance ($p < 0.05$) over the BEST MERGED model is indicated with [†]	EU-ES	BEST MERGED	46.2
		BEST MERGED + BT	47.1 [†]

Table 23 Comparison of results with baselines and best models

MODEL	ES-EU				EU-ES			
	TEST EITB		TEST ODE		TEST EITB		TEST ODE	
	BLEU	BP	BLEU	BP	BLEU	BP	BLEU	BP
ODE	28.6	1.000	32.4	0.934	34.0	0.980	46.7	0.953
STACC _{0.0}	34.7	0.781	11.3	0.594	44.8	0.910	24.5	0.861
LASER _{0.0}	36.4	0.835	12.0	0.664	44.9	0.925	25.1	0.881
ODE + STACC _{0.0}	36.1	0.812	30.6	0.936	46.5	0.939	44.7	0.957
ODE + LASER _{0.0}	36.9	0.849	30.2	0.921	46.8	0.951	45.0	0.950
BEST	39.6 [†]	1.000	30.5	0.947	47.1 [†]	0.958	44.4	0.957

In both translation directions, models trained on backtranslated data improved significantly over the ODE baseline, with gains of approximately 7 and 10 BLEU points for ES-EU and EU-ES, respectively. However, in both cases models based on aligned data performed significantly better, in particular for translation from Spanish to Basque. The model variants with tagged backtranslations improved marginally over the non-tagged version for ES-EU but slightly degraded the performance for EU-ES. These results are in line with those obtained in previous sections, where translation from Spanish to Basque required more filtering and data identification, whereas translation from Basque to Spanish seems to be optimal with the largest amount of available data.

As an additional experiment, we trained models based on a mixture of aligned and backtranslated data, where data discarded via the initial alignments, plus threshold and length filtering, were included as backtranslations. The main goal of this experiment was to evaluate the optimal combination of aligned and backtranslated data.

The results, shown in Table 22, indicate no improvement in terms of BLEU for ES-EU, but are impactful with respect to the length of the translated sentences, as indicated by results in terms of BP, with translations that are more similar in length to the human references overall. For EU-ES, the model trained on mixed data improved over the best merged model, with a statistically significant gain of 0.9

Table 24 Examples of aligned sentences in the final corpus, with English translations

LANG	EXAMPLE
ES	aún así , los partidos minoritarios buscan convertirse en la tercera fuerza política del país
EU	hala eta guztiz ere , gutxiengoen alderdiak herrialdeko hirugarren indar politikoa izateko borrokatzen dira
EN	<i>even so, minority parties seek to become the third political force in the country</i>
ES	Nassralla defiende que el productor alteró el contenido del film sin su conocimiento
EU	ekoizleak filmaren edukia bere oniritzia jaso gabe aldatu zuela adierazi du Nassrallak
EN	<i>Nassralla argues that the producer altered the content of the film without his knowledge</i>
ES	la sequía unida a las altas temperaturas , la baja humedad relativa y la intensidad del viento durante los próximos días hará que aumente la probabilidad de que se produzcan incendios forestales
EU	Lehorteek , temperatura altuek , hezetasun erlatibo baxuak eta haizearen intentsitateak datorren egunetan basoetan suteak izateko arriskua handitzea eragingo dute
EN	<i>Drought coupled with high temperatures , low relative humidity and wind intensity over the next few days will increase the likelihood of forest fires</i>
ES	el luxemburgués aceleró y se quedó atascado por la mecánica
EU	Luxemburgotarrak azeleratu egin zuen eta trabatuta geratu zen arazo mekanikoengatik
EN	<i>The Luxembourgian accelerated and got stuck because of mechanical issues</i>
ES	el 85 % de Nueva Orleans quedó bajo el agua , en algunas zonas a 7 metros de profundidad
EU	New Orleanseko lurraldearen % 85 urpean geratu zen , zenbait gunetan 7 metroko sakoneran
EN	<i>85 % of New Orleans was underwater , in some areas 7 meters deep</i>

BLEU points and translated sentences which were also closer to that of the human references as well, although not to the extent observed for translation from Spanish to Basque.

Overall, from the results on these datasets at least, backtranslations cannot be used as a replacement for comparable data alignment, but are a notable complement to reach optimal performance in terms of BLEU and to approximate reference translations in length.

11 Comparative summary

Starting from the baseline trained on the ODE corpus, various models have been trained and evaluated in the previous sections. Table 23 summarises the comparative results between the baselines and the best models achieved via data combination and filtering. For translation from Spanish to Basque, the best model was based on the union of the STACC_{0,0} and LASER_{1,0} datasets, length filtering of outliers, source data tagging and inclusion of filtered backtranslations. For translation from Basque to Spanish, the best model was based on STACC_{0,0} and LASER_{0,0} without length filtering or tagged data, along with backtranslations over the non-aligned data. The gains obtained by the best models were statistically significant over the baselines on the EITB test sets in all cases and both translation directions.

Table 24 provides some examples of aligned sentences in the final corpora. These examples illustrate the quality of the parallel resource obtained from the original comparable datasets, and the variety of topics covered in the corpus, including politics, world affairs, weather, sports and culture. The specific challenges presented by the morpho-syntactic properties of Basque, including agglutinative morphology, ergativity or relatively free word order, among others,¹¹ make it even more necessary to prepare additional parallel resources for this language. Comparable corpora can provide useful data to build such resources, although they require dedicated analyses and careful selection to fully exploit their potential.

12 Conclusions

In this work, we presented the results of a case study centred on the exploitation of Basque-Spanish comparable data for Neural Machine Translation. The original corpus is composed of news produced independently in these two languages by the Basque public broadcaster EITB. We applied and evaluated several techniques to create different variants of the corpus and measure their impact on machine translation quality.

Two efficient data alignment methods were evaluated, one based on lexical translation overlaps and the other on bilingual sentence embeddings. Both proved

¹¹ See Etchegoyhen et al. (2018) and references therein for more details on Machine Translation of Basque.

similarly efficient on their own to mine parallel data in the original corpus and train NMT models in both translation directions. The datasets extracted with each method were evaluated, and although they extracted a significant portion of similar data, manual analysis of the data showed that each method is able to mine valid parallel sentences that are not retrieved by the other approach. More work will thus be necessary to devise a single method with the combined accuracy of these two methods at least.

Different alignment thresholds were evaluated for each of the selected methods, with minor impact on translation quality on the models based on merged ODE and EITB datasets. Other studies, on this dataset and others, have shown the benefits of higher alignment thresholds with both methods, though this was not case for this particular experimental setup where the goal was to measure the impact of complementing parallel data with comparable data on NMT models.

Additionally, the impact of further filtering based on length-difference outliers was also measured, with the notable result that such filtering was necessary for Spanish to Basque translation, given information imbalance in the data, but not in the other translation direction, as NMT models proved able to benefit from target language information despite degenerate comparable source information. A related result was the tendency of NMT models to gear towards summarisation when provided with impoverished comparable target information, a phenomenon which is likely to arise with comparable corpora and needs to be controlled for an optimal exploitation of the data.

Results on the impact of tagging for comparable data were also discussed. This method was shown to be effective in helping the models discriminate noisy comparable data when there is less information in the target sentence than in the source, but detrimental in the opposite case, where the imbalanced comparable data may still strengthen target-side sequence modelling. Mixed tagging, where only lower quality alignments were signalled, resulted in lower quality overall, indicating that comparable data as a whole required identification in the training process, at least for the translation pair where tagging led to significant improvements.

Both length-based filtering and tagging were the most impactful methods overall, although the impact of these methods was dependent on the specific information imbalance in the dataset, with translation from Spanish to Basque as the scenario benefitting from these techniques for the comparable data used in our experiments.

Finally, we compared the use of aligned comparable sentences with simply exploiting backtranslations. Overall, comparable sentence alignment achieved higher translation quality, although complementary backtranslations proved useful to further increase the coverage of the translation models, via mixtures of aligned sentences and backtranslations. Comparable data may thus be exploited for Neural Machine Translation beyond the standard combination of alignment and threshold-based filtering.

Significant improvements in translation quality can be obtained with comparable corpora, and may be useful in particular for under-resourced languages. However, their use requires a careful analysis of phenomena such as information imbalance, in specific translation pairs, to fully exploit their potential for Neural Machine Translation.

Acknowledgements We wish to thank the Basque public broadcasting organisation EITB, for their support and their willingness to share the corpus with the community, and the anonymous LRE reviewers for their helpful comments. This work was partially supported the Department of Economic Development of the Basque Government (Spri), via projects ITAI (ZL-2021/00038) and TANDO (KK-2020/00074).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdul-Rauf, S., & 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* (pp. 16–23). Association for Computational Linguistics.
- Abdul-Rauf, S., & Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine Translation*, 25(4), 341–375.
- Artetxe, M., Labaka, G., Agirre, E., & Cho, K. (2018). Unsupervised neural machine translation. In *International Conference on Learning Representations*.
- Artetxe, M., & Schwenk, H. (2019). Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3197–3203). Association for Computational Linguistics.
- Azpeitia, A., & Etchegoyhen, T. (2019). Efficient document alignment across scenarios. *Machine Translation*, 33, 205–237.
- Azpeitia, A., Etchegoyhen, T., & Martínez Garcia, E. (2017). Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the Tenth Workshop on Building and Using Comparable Corpora* (pp. 41–45).
- Azpeitia, A., Etchegoyhen, T., & Martínez Garcia, E. (2018). Extracting parallel sentences from comparable corpora with STACC variants. In *Proceedings of the Eleventh Workshop on Building and Using Comparable Corpora* (pp. 48–52).
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.
- Belinkov, Y., & Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*.
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., & Turchi, M. (2017). Findings of the 2017 conference on machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers* (pp. 169–214).
- Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., & Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation* (pp. 131–198).
- Buck, C., & Koehn, P. (2016a). Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation* (pp. 554–563). Association for Computational Linguistics.
- Buck, C., & Koehn, P. (2016b). Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation* (pp. 672–678). Association for Computational Linguistics.

- Caswell, I., Chelba, C., & Grangier, D. (2019). Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)* (pp. 53–63). Association for Computational Linguistics.
- Chen, J., Chau, R., & Yeh, C.-H. (2004). Discovering parallel text from the World Wide Web. In *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation - Volume 32, ACSW Frontiers '04* (pp. 157–161). Australian Computer Society, Inc.
- Chen, J., & Nie, J.-Y. (2000). Parallel web text mining for cross-language IR. In *Content-Based Multimedia Information Access - Volume 1* (pp. 62–77). Centre des hautes études internationales d'informatique documentaire.
- Cheng, Y., Tu, Z., Meng, F., Zhai, J., & Liu, Y. (2018). Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1756–1766). Association for Computational Linguistics.
- Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of The 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Edurov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 489–500). Association for Computational Linguistics.
- Esplá-Gomis, M., Forcada, M. L., Ortiz-Rojas, S., & Ferrández-Tordera, J. (2016). Bitextor's participation in WMT'16: shared task on document alignment. In *Proceedings of the First Conference on Machine Translation* (pp. 685–691).
- Etchegoyhen, T., & Azpeitia, A. (2016a). A portable method for parallel and comparable document alignment. *Baltic Journal of Modern Computing*, 4(2), 243–255.
- Etchegoyhen, T., & Azpeitia, A. (2016b). Set-theoretic alignment for comparable corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers (pp. 2009–2018).
- Etchegoyhen, T., Azpeitia, A., & Pérez, N. (2016). Exploiting a large strongly comparable corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.
- Etchegoyhen, T., & Gete, H. (2020). Handle with care: A case study in comparable corpora exploitation for neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 3792–3800). European Language Resources Association.
- Etchegoyhen, T., Martínez García, E., Azpeitia, A., Labaka, G., Alegria, I., Cortes Etxabe, I., Jauregi Carrera, A., Ellakuria Santos, I., Martín, M., & Calonge, E. (2018). Neural machine translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 139–148).
- Fung, P., & 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* (pp. 57–63).
- Germann, U. (2016). Bilingual document alignment with latent semantic indexing. In *Proceedings of the First Conference on Machine Translation* (pp. 692–696). Association for Computational Linguistics.
- Gomes, L., & Lopes, G. P. (2016). First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation* (pp. 697–702). Association for Computational Linguistics.
- Grégoire, F., & Langlais, P. (2017). BUCC 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora* (pp. 46–50).
- Iglewicz, B., & Hoaglin, D. (1993). Volume 16: How to detect and handle outliers. *The ASQC basic references in quality control: statistical techniques*, 16.
- Ion, R., Ceașu, A., & 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* (pp. 128–135). Association for Computational Linguistics.
- Irvine, A., & Callison-Burch, C. (2013). Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation* (pp. 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.
- Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Hermann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., & Birch, A. (2018). Marian: Fast neural machine translation in C++. In *Proceedings of 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations* (pp. 116–121).
- Khayrallah, H., & Koehn, P. (2018). On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation* (pp. 74–83).
- Kobus, C., Crego, J., & Senellart, J. (2017). Domain control for neural machine translation. In *Proceedings of Recent Advances in Natural Language Processing* (pp. 372–378). INCOMA Ltd.
- Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 388–395).
- 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 Association for Computational Linguistics* (pp. 177–180).
- Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.
- Li, B., & Gaussier, E. (2013). Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In *Building and Using Comparable Corpora* (pp. 131–149).
- Ma, X., & Liberman, M. (1999). Bits: A method for bilingual text search over the web. In *Proceedings of Machine Translation Summit VII* (pp. 538–542).
- Morin, E., Hazem, A., Boudin, F., & Clouet, E. L. (2015). Lina: Identifying comparable documents from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*.
- Munteanu, D. S., & Marcu, D. (2002). Processing comparable corpora with bilingual suffix trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 289–295).
- Munteanu, D. S., & Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4), 477–504.
- Papavassiliou, V., Prokopidis, P., & Piperidis, S. (2016). The ILSP/ARC submission to the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation* (pp. 733–739).
- Papineni, K., Roukos, S., Ward, T., & 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* (pp. 311–318).
- Poncelas, A., Shterionov, D. S., Way, A., de Buy Wenniger, G. M., & Passban, P. (2018). Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 249–258).
- Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191).
- Rapp, R. (1995). Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 320–322).
- Resnik, P. (1999). Mining the Web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (pp. 527–534).
- Resnik, P., & Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3), 349–380.
- Sarikaya, R., Maskey, S., Zhang, R., Jan, E.-E., Wang, D., Ramabhadran, B., & Roukos, S. (2009). Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. In *Proceedings of InterSpeech* (pp. 432–435).
- Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 228–234).
- Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). WikiMatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. *CoRR*, abs/1907.05791.
- Sennrich, R., Haddow, B., & Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the*

- Association for Computational Linguistics: Human Language Technologies* (pp. 35–40). Association for Computational Linguistics.
- Sennrich, R., Haddow, B., & Birch, A. (2016b). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 86–96).
- Sennrich, R., Haddow, B., & Birch, A. (2016c). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (pp. 1715–1725).
- Sharoff, S., Rapp, R., Zweigenbaum, P., & Fung, P. (2016). *Building and using comparable corpora*. Incorporated: Springer Publishing Company.
- Sharoff, S., Zweigenbaum, P., & Rapp, R. (2015). BUCC shared task: Cross-language document similarity. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora* (pp. 74–78).
- Skadipa, I., Aker, A., Mastropavlos, N., Su, F., Tufis, D., Verlic, M., Vasiljevs, 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*.
- Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 403–411).
- Sperber, M., Niehues, J., & Waibel, A. (2017). Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Workshop on Spoken Language Translation* (pp. 90–96).
- Stefănescu, D., Ion, R., & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th Conference of the European Association for Machine Translation* (pp. 137–144).
- Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th Language Resources and Evaluation Conference* (pp. 2214–2218).
- Uszkoreit, J., Ponte, J. M., Popat, A. C., & Dubiner, M. (2010). Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1101–1109).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 5998–6008).
- Yamagishi, H., Kanouchi, S., Sato, T., & Komachi, M. (2016). Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation* (pp. 203–210).
- Zafarian, A., Aghasadeghi, A., Azadi, F., Ghiasifard, S., Alipanahloo, Z., Bakhshaei, S., & Ziabary, S. M. M. (2015). AUT document alignment framework for bucc workshop shared task. In *Proceedings of the 8th Workshop on Building and Using Comparable Corpora* (p. 79).
- Zhao, B., & Vogel, S. (2002). Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining* (pp. 745–748).
- Zweigenbaum, P., Sharoff, S., & Rapp, R. (2017). Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora* (pp. 60–67). Association for Computational Linguistics.
- Zweigenbaum, P., Sharoff, S., & Rapp, R. (2018). Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In Rapp, R., Zweigenbaum, P., and Sharoff, S. (Eds.) *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.