Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3799–3807
Marseille, 11–16 May 2020

Handle with Care: A Case Study in Comparable Corpora Exploitation for Neural Machine Translation

Thierry Etchegoyhen and Harritxu Gete Ugarte
Department of Speech and Natural Language Technologies, Vicomtech
Mikeletegi Pasalekua, 57, Donostia, Gipuzkoa, Spain
{tetchegoyhen, hgete}@vicomtech.org

Abstract

We present the results of a case study in the exploitation of comparable corpora for Neural Machine Translation. A large comparable corpus for Basque-Spanish was prepared, on the basis of independently-produced news by the Basque public broadcaster EITB, and we discuss the impact of various techniques to exploit the original data in order to determine optimal variants of the corpus. In particular, we show that filtering in terms of alignment thresholds and length-difference outliers has a significant impact on translation quality. The impact of tags identifying comparable data in the training datasets is also evaluated, with results indicating that this technique might be useful to help the models discriminate noisy information, in the form of informational imbalance between aligned sentences. The final corpus was prepared according to the experimental results and is made available to the scientific community for research purposes.

Keywords: Comparable Corpora, Basque, Spanish, Neural Machine Translation

1. Introduction

Comparable corpora are an important source of potential parallel data, suitable to train data-driven machine translation systems (Munteanu and Marcu, 2005; Sharoff et al., 2016) or to create bilingual dictionaries (Rapp, 1995).

The extraction of parallel sentences from this type of corpora faces a number of challenges, since potential parallel data are immersed in vast amounts of unrelated content which need to be efficiently mined. Over the years, several techniques have been designed to address comparable document alignment (Sharoff et al., 2015; Azpeitia and Etchegoyhen, 2019) and comparable sentence alignment (Munteanu and Marcu, 2002; Fung and Cheung, 2004; Smith et al., 2010; Etchegoyhen and Azpeitia, 2016b; Artetxe and Schwenk, 2019), 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; Bojar et al., 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 and Bisk, 2018; Sperber et al., 2017; Cheng et al., 2018; Khayrallah and Koehn, 2018). 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 preparation of a large comparable corpus for Basque-Spanish, composed of independently-produced news by the Basque public broadcaster EITB[1], and focus on the impact of various techniques to exploit the original data for Neural Machine Translation. We show in particular that filtering in terms of alignment thresholds and length-difference outliers have a significant impact on translation quality. The impact of tags identifying comparable data in the training datasets is also evaluated, following the approach proposed by Caswell et al. (2019) for synthetic data, with results indicating that this technique might be useful 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.

The final corpus is available to the community for research purposes, and the main contributions of our work are thus the following:

- An evaluation of various techniques to exploit comparable corpora for Neural Machine Translation.
- An evaluation of data tagging for comparable corpora, which, to our knowledge, has not been explored yet.
- A large parallel corpus in the news domain for Basque-Spanish, an under-resourced language pair, shared with the scientific community.

The remainder of this paper is organised as follows: Section 2. describes related work on comparable corpora; Section 3. describes the corpora used in this study; in Section 4., we describe the various experiments performed with the comparable and parallel Basque-Spanish corpora: the experimental setup is described in Section 4.1., Section 4.2. discusses the results obtained with different sentence alignment thresholds, Section 4.3. centres on the results obtained with length filtering, Section 4.4. describes our tagging approach and its results, and Section 4.5. summarises the characteristics of the final corpus, obtained by combining the optimal methods determined in the previous sections; finally, Section 5. draws conclusions from this work.

2. Related work

A significant body of work has been produced over the years to mine parallel sentences from large collections of

[1] 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	3,745,794	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

Table 1: Original EITB corpus 2009-2018

monolingual corpora (Sharoff et al., 2016), 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 (Resnik and Smith, 2003; Chen and Nie, 2000; Munteanu and Marcu, 2005; Papavassiliou et al., 2016). 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 and Liberman, 1999), mutual information (Fung and Cheung, 2004), expectation-maximisation (Ion et al., 2011) or n-gram matching using machine-translated documents (Uszkoreit et al., 2010). Several approaches have used a mixture of content and structural properties, notably the systems in the WMT 2016 document alignment shared task (Buck and 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 and 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 and Marcu, 2002), maximum likelihood (Zhao and Vogel, 2002), binary classification (Munteanu and Marcu, 2005), cosine similarity (Fung and Cheung, 2004), reference metrics over statistical machine translations (Abdul-Rauf and Schwenk, 2009; Sarikaya et al., 2009) or rich features (Stefănescu et al., 2012; Smith et al., 2010). In recent years deep learning approaches have been applied to the task as well, with bidirectional recurrent neural networks (Grégoire and 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 as well (Etchegoyhen and Azpeitia, 2016b), improving over more sophisticated feature-rich methods. This approach, 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; Pierre Zweigenbaum and Rapp, 2018). Improvement over these results were obtained with the margin-based approach of Artetxe and Schwenk (2019), which uses cosine similarity over multilingual sentence embeddings, extended with a filtering mechanism based on nearest neighbours similarity.

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 what follows, we exploit a larger version of the corpus, built with news generated in subsequent years, and evaluate the impact of different methods to produce a parallel corpus from the comparable data.

3. Corpora

The EITB corpus is composed of news independently produced in Basque and Spanish by the journalists of the Basque Country's public broadcast service.[2], to report on the same specific events. The corpus can thus be considered strongly comparable (Skadiņa et al., 2012) and viewed as a rich source of parallel data for this language pair (Etchegoyhen et al., 2016).

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

This version of the original dataset covers ten years of content, and is composed of $168,984$ documents in Basque and $174,348$ in Spanish, extracted from XML files with the structure described in Etchegoyhen et al. (2018b). With an original amount of over two million sentences per language, it is the largest available corpus 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.[3]. We used the corpus created from these translation memories and shared by Etchegoyhen et al. (2018a),[4] to which we will refer as OPENDATA in what follows. The corpus is constituted mainly by translations of administrative content, notably from the Instituto Vasco de Administración Pública (IVAP). As it consists of professionally produced translations, the OPENDATA corpus was used as a parallel basis for the experiments described in what follows.

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 STACC system (Etchegoyhen and Azpeitia, 2016b), in its variant with lexical weighting (Azpeitia et al., 2017). To extract the translation tables necessary in these approaches to document and sentence alignment, we used the FASTALIGN toolkit (Dyer et al., 2013) on the OPENDATA 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 OPENDATA and EITB corpora, in number of aligned sentences are shown in Table 9, where the latter was obtained after enforcing one-to-one alignment but without any filtering based on the alignment scores computed by STACC, i.e. with an alignment threshold set to 0.

CORPUS	OPENDATA	$\text{EITB}_{0.0}$
TRAIN	921,763	1,337,040
DEV	2,000	2,000
TEST	-	1,678

Table 2: Corpora statistics (number of sentence pairs)

The development and test sets were extracted from the EITB dataset, with manually verified alignments. For the test set, additional Basque translations were manually created by professional translators, to increase the robustness of the test considering the relatively free word order in Basque (Etchegoyhen et al., 2018a). In the Spanish to Basque translation direction, we thus 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.

[3] https://opendata.euskadi.eus

[4] In the version shared for the IWSLT 2018 shared task, available at: https://sites.google.com/site/iwsltevaluation2018/TED-tasks

4. Experiments

In this Section, we first describe the experimental setup, including initial translation models. We then describe the different methods in turn and discuss their results.

4.1. Experimental Setup

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 and 8 attention heads, reproducing the basic transformer described in the original paper. 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 for a given sentence length that fits the specified memory. The validation data was evaluated every 3,500 steps, and the training process ended if there was no improvement in the perplexity of 5 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 two initial models, on the OPENDATA corpus and on that same corpus merged with the aligned EITB corpus with an alignment threshold set to 0, to measure the initial contribution of the selected comparable data. 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 two initial models are shown in Table 3.

In both translation directions, the contribution of the comparable data was significant, with large improvements in BLEU scores. Although this is not unexpected, since the OPENDATA 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).

MODEL	ES-EU	EU-ES
OPENDATA	28.6	34.0
$\text{OPENDATA+EITB}_{0.0}$	36.1	46.5

Table 3: BLEU results with initial models

4.2. Alignment Thresholds

Methods to extract parallel data from comparable corpora rely on metrics that measure translation equivalence in some forms. 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 (Etchegoyhen and Azpeitia, 2016b; Artetxe and Schwenk, 2019). 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 different subsets of the aligned EITB corpus after applying

different thresholds, selected according to the ranges of the STACC metric which produced significant amounts of data. Each subset was then used to fine-tune the model based on the OPENDATA corpus, with the results shown in Table 4 in terms of BLEU score and number of aligned sentences.

THRESHOLD	ALIGNED	BLEU	
		ES-EU	EU-ES
0.0	1,337,040	34.8	44.3
0.15	1,122,890	35.3	**44.5**
0.17	930,839	**36.1**	44.2
0.20	580,478	35.6	43.8

Table 4: Alignment threshold results

As these results demonstrate, selecting subsets extracted with higher alignment thresholds is beneficial overall, although the loss of data incurred with more restrictive thresholds can be sub-optimal. For the experiments in the next sections, we selected the dataset extracted with a threshold of 0.17, as it provided the optimal balance in terms of BLEU for the two translation directions.[5]

4.3. 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. (2018a), 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 OPENDATA 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 corpus with threshold 0, according to the formula in Equation 1:

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

, where x is the length difference of a sentence pair in the EITB corpus, $\tilde{y}$ 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 in the aligned EITB corpus with threshold 0.17, selected after the results in the previous section. 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.[6]

The results on fine-tuned models trained on each selected sub-corpus filtered by length outliers are shown in Table 5. Also indicated in this table are the size of the filtered corpus, the BLEU brevity penalty (BP), and the proportion of sentences where the length of the Spanish sentence is larger than that of the Basque sentence.

For Spanish to Basque translation, in terms of BLEU scores, length filtering improved over the unfiltered corpus, showing that information imbalance was significantly detrimental. For Basque to Spanish, the results were reversed, with a gradual decrease of BLEU scores with additional filtering. One interpretation of these 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, where for instance determiners are suffixes, more aggressive filtering of length-difference outliers lowers the proportion of Spanish sentences that are longer than their Basque counterparts, 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 the latter language to the former would thus have the effect of orienting the models towards summarisation, with an impact on translation quality that needs to be compensated with more length-based filtering. This conjecture is supported by the 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 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 its modelling of the target sequences in the face of degenerate source input.

Selecting a single corpus based on these results faces a difficult choice. Although the Basque to Spanish results tend to indicate that the non-filtered output provides significant improvements in terms of BLEU, for the reasons hypothesised above, the opposite translation direction would favour the selection of a corpus based on length filtering with a 2.0 outliers threshold, as it provides the best BLEU score and lower impact in terms of brevity than a 3.5 threshold. In what follows, we will select the corpus based on the 2.0 threshold,

[5] As a side note, the results obtained via fine-tuning (Luong and Manning, 2015) are lower than those obtained via training over merged datasets, as shown by the results in Table 3 compared to those in Table 4 with threshold 0. This is not unexpected (Crego et al., 2016) and does not impact the relative comparisons established between the fine-tuned models using different thresholds.

[6] 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.

CORPUS	ALIGNED	ES-EU		EU-ES		len(filt$_{es}$) > len(filt$_{eu}$)
		BLEU	BP	BLEU	BP	
EITB$_{0.17}$	930,839	36.1	0.811	44.2	0.919	-
EITB$_{0.17_lgs3.5}$	773,755	37.7	0.916	43.6	0.919	98.84%
EITB$_{0.17_lgs2.0}$	637,183	37.7	0.960	42.2	0.907	98.65%
EITB$_{0.17_lgs1.5}$	580,448	37.0	0.979	41.8	0.900	98.48%

Table 5: Fine-tuning results with length filtering thresholds

LANG	EXAMPLE
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 hizkuntzatara itzultzen da , txinerara tartean Txinako merkatu berriak barne hartzearren .
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* .
EU	bat baino gehiago emanaldi amaierara iristen bada lehiaketaren bigarren epea jarriko da martxan .

Table 6: Examples of filtered sentence pairs with modified z-score above 2.0

MODEL	ES-EU				EU-ES			
	TAG		NO-TAG		TAG		NO-TAG	
	BLEU	BP	BLEU	BP	BLEU	BP	BLEU	BP
OPENDATA	-	-	28.6	1.0	-	-	34.0	0.98
OPENDATA+EITB$_{0.0}$	38.6	0.984	36.1	0.812	45.4	0.970	46.5	0.939
OPENDATA+EITB$_{0.17}$	39.5	0.989	35.7	0.828	44.7	0.965	46.1	0.943
OPENDATA+EITB$_{0.17_lgs2.0}$	38.2	0.992	38.9	0.973	43.4	0.970	44.1	0.927

Table 7: Results on merged datasets with and without tags

to favour the translation direction with lower results overall, noting that the filtered sentence pairs could be used to train a separate model for Basque to Spanish only. Selecting this filtering threshold also improves the overall quality of the corpus to be shared, as it removes imbalanced pairs of the type shown in Table 6, where the information in the Spanish sentence that is missing in the Basque counterpart of the aligned pair is marked in italics.

4.4. 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 for back-translated synthetic data, among others.

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 OPENDATA corpus with selected variants of the EITB corpus, with and without tags indicating comparable data. The results of these experiments are shown in Table 7.

For Spanish to Basque, tagging was only effective on the noisier datasets, i.e. the EITB variant with no filtering of length-difference outliers. For the less noisy dataset, based on a higher alignment threshold and 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 0.812 for the untagged model based on the EITB$_{0.0}$ model, to 0.984 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 the natural translations, as in the heavily filtered variants.

For Basque to Spanish, tagging was detrimental overall in terms of BLEU, despite minor improvements regarding the brevity of translations. The tendency was the same across dataset variants, irrespective of filtering. 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 could be further examined by comparing tagged back-translated data with tagged comparable data, an exploration we leave for future research.

LANG	EXAMPLE
ES	aún$_{1,2}$ así$_{1,2}$, los$_4$ partidos$_4$ minoritarios$_3$ buscan$_5$ convertirse$_{6,7}$ en$_{6,7}$ la$_8$ tercera$_9$ fuerza$_{11}$ política$_{10}$ del$_{12,13}$ país$_{14}$.
EU	hala$_{1,2}$ eta$_{1,2}$ guztiz$_{1,2}$ ere$_{1,2}$, gutxiengoen$_3$ alderdiak$_4$ herrialdeko$_{12,13,14}$ hirugarren$_{8,9}$ indar$_{11}$ politikoa$_{10}$ izateko$_{6,7}$ borrokatzen$_5$ dira$_5$.
EN	*even$_1$ so$_2$, minority$_3$ parties$_4$ seek$_5$ to$_6$ become$_7$ the$_8$ third$_9$ political$_{10}$ force$_{11}$ in$_{12}$ the$_{13}$ country$_{14}$* .
ES	Nassralla$_1$ defiende$_2$ que$_3$ el$_4$ productor$_5$ alteró$_6$ el$_7$ contenido$_8$ del$_{9,10}$ film$_{11}$ sin$_{12}$ su$_{13}$ conocimiento$_{14}$.
EU	ekoizleak$_{4,5}$ filmaren$_{9,10,11}$ edukia$_{7,8}$ bere$_{13}$ oniritzia$_{14}$ jaso$_{12}$ gabe$_{12}$ aldatu$_6$ zuela$_{3,6}$ adierazi$_2$ du$_2$ Nassrallak$_1$
EN	*Nassralla$_1$ argues$_2$ that$_3$ the$_4$ producer$_5$ altered$_6$ the$_7$ content$_8$ of$_9$ the$_{10}$ film$_{11}$ without$_{12}$ his$_{13}$ knowledge$_{14}$* .
ES	la$_1$ sequía$_1$ unida$_2$ a$_3$ las$_5$ altas$_4$ temperaturas$_5$, la$_8$ baja$_6$ humedad$_8$ relativa$_7$ y$_9$ la$_{11}$ intensidad$_{11}$ del$_{10,11}$ viento$_{10}$ durante$_{12}$ los$_{13}$ próximos$_{14,15}$ días$_{16}$ hará$_{17}$ que$_{17}$ aumente$_{18}$ la$_{19}$ probabilidad$_{20}$ de$_{21}$ que$_{20,21}$ se$_{20,21}$ produzcan$_{20,21}$ incendios$_{23}$ forestales$_{22}$.
EU	Lehorteek$_1$, tenperatura$_5$ altuek$_4$, hezetasun$_8$ erlatibo$_7$ baxuak$_6$ eta$_9$ haizearen$_{10}$ intentsitateak$_{11}$ datorren$_{13,14,15}$ egunetan$_{12,16}$ basoetan$_{22}$ suteak$_{23}$ izateko$_{20}$ arriskua$_{20}$ handitzea$_{18}$ eragingo$_{17,18}$ dute$_{18}$.
EN	*drought$_1$ coupled$_2$ with$_3$ high$_4$ temperatures$_5$, low$_6$ relative$_7$ humidity$_8$ and$_9$ wind$_{10}$ intensity$_{11}$ over$_{12}$ the$_{13}$ next$_{14}$ few$_{15}$ days$_{16}$ will$_{17}$ increase$_{18}$ the$_{19}$ likelihood$_{20}$ of$_{21}$ forest$_{22}$ fires$_{23}$* .
ES	el$_1$ luxemburgués$_2$ aceleró$_3$ y$_4$ se$_5$ quedó$_5$ atascado$_6$ por$_7$ la$_8$ mecánica$_8$.
EU	Luxenburgotarrak$_{1,2}$ azeleratu$_3$ egin$_3$ zuen$_3$ eta$_4$ trabatuta$_6$ geratu$_5$ zen$_5$ arazo$_7$ mekanikoengatik$_{7,8}$.
EN	*the$_1$ Luxembourgian$_2$ accelerated$_3$ and$_4$ got$_5$ stuck$_6$ by$_7$ mechanics$_8$* .
ES	el$_1$ 85$_1$ %$_2$ de$_3$ Nueva$_4$ Orleáns$_5$ quedó$_6$ bajo$_7$ el$_7$ agua$_7$, en$_8$ algunas$_9$ zonas$_{10}$ a$_{13}$ 7$_{11}$ metros$_{12}$ de$_{13}$ profundidad$_{13}$.
EU	New$_4$ Orleanseko$_5$ lurraldearen$_{4,5}$ %$_2$ 85$_1$ urpean$_7$ geratu$_6$ zen$_6$, zenbait$_9$ gunetan$_{8,10}$ 7$_{11}$ metrotako$_{12}$ sakoneran$_{13}$.
EN	*85$_1$ %$_2$ of$_3$ New$_4$ Orleans$_5$ was$_6$ underwater$_7$, in$_8$ some$_9$ areas$_{10}$ 7$_{11}$ meters$_{12}$ deep$_{13}$* .

Table 8: Examples of aligned sentences in the final corpus, with English translations

4.5. Shared Corpus

In view of the results discussed in the previous sections, the optimal variants of the datasets in terms of BLEU scores of the resulting models are different depending on the translation direction. For Basque to Spanish, the corpus without either alignment or length-difference filtering would provide the best objective scores, whereas from Spanish to Basque the variant with an alignment threshold of 0.17, augmented with tags, would lead to the best results in terms of translation metric scores.

Our aim in sharing a prepared dataset was however to select a unique parallel corpus with the highest alignment quality which could be beneficial in both translation directions, without tagging nor the indirect BLEU improvements obtained from the ability of NMT models to exploit degenerate information as input. We therefore selected the variant of the EITB corpus based on an alignment threshold of 0.17 and length-based filtering with a threshold of 2.0, as it complied with this requirement. With 637.183 aligned sentences, it will be the largest parallel corpora for Basque-Spanish in the news domain, covering a large amount of topics.

Table 8 provides some examples of aligned sentences in the final corpus.[7] 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.

[7] To facilitate the understanding of the Basque and Spanish examples, we include a likely English translation in each case, along with indexes to indicate approximate word correspondences between the English sentence, where each token (excluding punctuation) is assigned a separate index, and the Basque and Spanish sentences.

The specific challenges presented by the morpho-syntactic properties of Basque, including agglutinative morphology, ergativity or relatively free word order, among others,[8] make it even more necessary to prepare additional parallel resources for this language. We hope that the corpus prepared and described in this work will help advance research and development for the under-resourced Basque-Spanish language pair.

The corpus will be made available on OPUS[9] (Tiedemann, 2012) under the Creative Commons CC-BY-NC-SA license. The statistics for the shared version of the corpus are shown in Table 9.

	ES	EU
SENTENCES	637,183	637,183
WORDS	13,365,220	10,882,709

Table 9: Shared EITB corpus statistics

5. Conclusions

In this article, we described the results of a case study on the exploitation of a large corpus of strongly comparable data for Basque-Spanish. The corpus is composed of news produced independently in these two languages by the Basque public broadcaster EITB and is shared with the community for scientific purposes.

We applied and evaluated several techniques to exploit the comparable data for Neural Machine Translation. Different alignment thresholds were thus evaluated, leading to the

[8] See (Etchegoyhen et al., 2018a) and references therein for more details on Machine Translation of Basque.

[9] http://opus.nlpl.eu/

selection of an optimal subset which could serve models in both translation directions. Additionally, the impact of further filtering based on length-difference outliers was also measured, with the notable result that such filtering is necessary for Spanish to Basque translation, given information imbalance in the data, but not in the other translation directions, as NMT models proved able to benefit from target language information despite degenerate comparable source information.

Another result of this study was the tendency of NMT models to gear towards summarisation when provided with impoverished 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 presented, a topic which had not been previously studied, to the best of our knowledge. Tagging was shown to be effective only in helping the models discriminate noisy comparable data, identified to be data with degenerate information in the target language. In all other cases, with either filtered datasets or higher informational content in the target side, tagging was detrimental and further experiments will be necessary, in particular to determine the precise types of comparable data which constitute noise for NMT models.

The usefulness of comparable corpora, in particular for under-resourced languages, has been comforted in this work, with large improvements in translation quality resulting from their use. Additionally, our results indicate that this type of data needs to be handled with care to benefit Neural Machine Translation approaches.

6. Acknowledgements

This work was supported by the Department of Economic Development and Competitiveness of the Basque Government, via the MODENA and COMETA projects. We wish to thank the Basque public broadcasting organisation EITB for their support and their willingness to share the corpus with the community.

7. Bibliographical 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.

Artetxe, M. and 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, pages 3197–3203, Florence, Italy, July. Association for Computational Linguistics.

Azpeitia, A. and Etchegoyhen, T. (2019). Efficient Document Alignment Across Scenarios. *Machine Translation*, 33:205–237.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2017). Weighted Set-Theoretic Alignment of Comparable Sentences. In Proceedings of the Tenth Workshop on Building and Using Comparable Corpora, pages 41–45.

Azpeitia, A., Etchegoyhen, T., and Martínez Garcia, E. (2018). Extracting Parallel Sentences from Comparable Corpora with STACC Variants. In Reinhard Rapp, et al., editors, Proceedings of the Eleventh Workshop on Building and Using Comparable Corpora, pages 48–52.

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

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.

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., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In Proceedings of the First Conference on Machine Translation, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

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., and Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT17). In Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers, pages 169–214. Association for Computational Linguistics, September.

Buck, C. and Koehn, P. (2016a). Findings of the wmt 2016 bilingual document alignment shared task. In Proceedings of the First Conference on Machine Translation, pages 554–563, Berlin, Germany.

Buck, C. and Koehn, P. (2016b). Quick and reliable document alignment via tf/idf-weighted cosine distance. In Proceedings of the First Conference on Machine Translation, pages 672–678.

Caswell, I., Chelba, C., and Grangier, D. (2019). Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 53–63, Florence, Italy, August. Association for Computational Linguistics.

Chen, J. and Nie, J.-Y. (2000). Parallel Web Text Mining for Cross-language IR. In Content-Based Multimedia Information Access - Volume 1, RIAO '00, pages 62–77, Paris, France, France. Centre des hautes études internationales d'informatique documentaire.

Chen, J., Chau, R., and 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, pages 157–161, Darlinghurst, Australia, Australia. Australian Computer Society, Inc.

Cheng, Y., Tu, Z., Meng, F., Zhai, J., and Liu, Y. (2018). Towards robust neural machine translation. In Proceedings of the 56th Annual Meeting of the Association for Com-

putational Linguistics (Volume 1: Long Papers), pages 1756–1766, Melbourne, Australia, July. Association for Computational Linguistics.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., et al. (2016). Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Dyer, C., Chahuneau, V., and 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.

Esplá-Gomis, M., Forcada, M. L., Ortiz-Rojas, S., and Ferrández-Tordera, J. (2016). Bitextor's participation in wmt'16: shared task on document alignment. In Proceedings of the First Conference on Machine Translation, pages 685–691.

Etchegoyhen, T. and Azpeitia, A. (2016a). A Portable Method for Parallel and Comparable Document Alignment. *Baltic Journal of Modern Computing*, 4(2):243–255. *Special Issue: Proceedings of EAMT 2016*.

Etchegoyhen, T. and 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, pages 2009–2018.

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

Etchegoyhen, T., Martínez Garcia, E., Azpeitia, A., Labaka, G., Alegria, I., Cortes Etxabe, I., Jauregi Carrera, A., Ellakuria Santos, I., Martin, M., and Calonge, E. (2018a). Neural Machine Translation of Basque. In Proceedings of the 21st Annual Conference of the European Association for Machine Translation, pages 139–148.

Etchegoyhen, T., Torné, A. F., Azpeitia, A., Garcia, E. M., and Matamala, A. (2018b). Evaluating Domain Adaptation for Machine Translation Across Scenarios. In Proceedings of the 11th Edition of the Language Resources and Evaluation Conference.

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

Germann, U. (2016). Bilingual document alignment with latent semantic indexing. pages 692–696.

Gomes, L. and Lopes, G. P. (2016). First steps towards coverage-based document alignment. In Proceedings of the First Conference on Machine Translation, pages 697–702.

Grégoire, F. and Langlais, P. (2017). BUCC 2017 Shared Task: a First Attempt Toward a Deep Learning Framework for Identifying Parallel Sentences in Comparable Corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 46–50. Association for Computational Linguistics.

Iglewicz, B. and Hoaglin, D. (1993). Volume 16: how to detect and handle outliers. *The ASQC basic references in quality control: statistical techniques*, 16.

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.

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., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast neural machine translation in C++. In Proceedings of ACL 2018, System Demonstrations, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Khayrallah, H. and Koehn, P. (2018). On the impact of various types of noise on neural machine translation. *CoRR*, abs/1805.12282.

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pages 372–378, Varna, Bulgaria, September. INCOMA Ltd.

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.

Li, B. and Gaussier, E. (2013). Exploiting comparable corpora for lexicon extraction: Measuring and improving corpus quality. In *Building and Using Comparable Corpora*. Springer, pp. 131–149.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In Proceedings of the International Workshop on Spoken Language Translation.

Ma, X. and Liberman, M. (1999). Bits: A method for bilingual text search over the web. In Machine Translation Summit VII, pages 538–542.

Morin, E., Hazem, A., Boudin, F., and Clouet, E. L. (2015). Lina: Identifying comparable documents from Wikipedia. In Eighth Workshop on Building and Using Comparable Corpora.

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

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

Papavassiliou, V., Prokopidis, P., and Piperidis, S. (2016). The ilsp/arc submission to the wmt 2016 bilingual docu-

ment alignment shared task. In Proceedings of the First Conference on Machine Translation, pages 733–739.

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, pages 311–318. Association for Computational Linguistics.

Pierre Zweigenbaum, S. S. and Rapp, R. (2018). Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Reinhard Rapp, et al., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France, may. European Language Resources Association (ELRA).

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

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Resnik, P. (1999). Mining the web for bilingual text. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pages 527–534, College Park, Maryland, USA, June. Association for Computational Linguistics.

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

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

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), pages 228–234. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and 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, pages 35–40, San Diego, California, June. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and 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), pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and 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, pages 1715–1725.

Sharoff, S., Zweigenbaum, P., and Rapp, R. (2015). Bucc shared task: Cross-language document similarity. *Proceedings of the 8th Workshop on Building and Using Comparable Corpora*, pages 74–78.

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

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

Sperber, M., Niehues, J., and Waibel, A. (2017). Toward robust neural machine translation for noisy input sequences. In International Workshop on Spoken Language Translation (IWSLT).

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.

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

Uszkoreit, J., Ponte, J. M., Popat, A. C., and Dubiner, M. (2010). Large scale parallel document mining for machine translation. In Proceedings of the 23rd International Conference on Computational Linguistics, pages 1101–1109. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, pages 6000–6010.

Yamagishi, H., Kanouchi, S., Sato, T., and 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 (WAT2016), pages 203–210, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Zafarian, A., Aghasadeghi, A., Azadi, F., Ghiasifard, S., Alipanahloo, Z., Bakhshaei, S., and Ziabary, S. M. M. (2015). Aut document alignment framework for bucc workshop shared task. *ACL-IJCNLP 2015*, page 79.

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

Zweigenbaum, P., Sharoff, S., and Rapp, R. (2017). Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 60–67. Association for Computational Linguistics.