Neural Machine Translation of Basque

Thierry Etchegoyhen,[1] **Eva Martínez Garcia,**[1] **Andoni Azpeitia,**[1]
Gorka Labaka,[2] **Iñaki Alegria,**[2] **Itziar Cortes Etxabe,**[3] **Amaia Jauregi Carrera,**[3]
Igor Ellakuria Santos,[4] **Maite Martin,**[5] **Eusebi Calonge**[5]
[1]Vicomtech - {tetchegoyhen, emartinez, aazpeitia}@vicomtech.org
[2]IXA taldea, University of the Basque Country - {i.alegria, gorka.labaka}@ehu.eus
[3]Elhuyar - {i.cortes, a.jauregi}@elhuyar.eus
[4]ISEA - isantos@iseamcc.net
[5]Ametzagaiña - maite@adur.com, ecalonge@ametza.com

Abstract

We describe the first experimental results in neural machine translation for Basque. As a synthetic language featuring agglutinative morphology, an extended case system, complex verbal morphology and relatively free word order, Basque presents a large number of challenging characteristics for machine translation in general, and for data-driven approaches such as attention-based encoder-decoder models in particular. We present our results on a large range of experiments in Basque-Spanish translation, comparing several neural machine translation system variants with both rule-based and statistical machine translation systems. We demonstrate that significant gains can be obtained with a neural network approach for this challenging language pair, and describe optimal configurations in terms of word segmentation and decoding parameters, measured against test sets that feature multiple references to account for word order variability.

1 Introduction

Neural machine translation (NMT) is fast becoming the dominant paradigm in both academic research and commercial exploitation, as evidenced in particular by large machine translation providers turning to NMT for their production engines (Crego et al., 2016; Wu et al., 2016) and NMT systems achieving the best results in most cases on standard shared tasks datasets (Bojar, 2016).

Sequence-to-sequence neural networks have proved effective in modelling translation phenomena (Sutskever et al., 2014). In particular, attentional encoder-decoder models (Bahdanau et al., 2015) have become a default NMT architecture, with other architectural variants explored in recent work (Vaswani et al., 2017; Gehring et al., 2016). These models have already been applied to a wide range of languages, initially on the most studied European languages and recently to a larger set of cases that includes morphologically rich languages (Bojar, 2017).

In this article we explore the applicability of neural machine translation to Basque, a language with noteworthy characteristics that may represent a challenge for encoder-decoder approaches with attention mechanisms.

First, Basque is a synthetic language that features agglutinative morphology, i.e. where words can be formed via morphemic sequences, and a large number of case affixes that mark ergativity, datives, different types of locatives and genitives, instrumentality, comitativity or causality, among others. Verbal morphology is also relatively rare, displaying complex forms that include subject, direct object, indirect object and allocutive agreement markers, with number, tense and aspect being marked as well. This kind of rich morphology raises difficulties in terms of word representations and drastically increases data sparseness issues. A detailed description of Basque grammar can be found in (De Rijk and De Coene, 2008).

Secondly, although phrases in this language present a rather fixed inner order, as exemplified for instance by the regular structure of noun phrases,[1] at the sentential level the ordering is rela-

[1] Although regular, the structure of noun phrases may also be challenging, with left-branching relative clauses and affixa-

tively free. Syntactically, order is essentially determined in terms of focus and topic. Although different orderings mostly reflect underlying variations according to these notions, for translation between Basque and languages with more rigid syntax the end-result is higher variability in terms of sentential input and output. Such variations may represent an additional challenge for NMT models that manage input information via learned attentional distributions and generate translations via decoding processes based on the previously generated element and beam searches.

Finally, Basque is a low-resourced language, with few publicly available parallel corpora. This is a third challenge for data-driven approaches in general, and NMT in particular as it usually requires larger training resources than statistical machine translation (Zoph et al., 2016).

To explore these challenges, we built several large neural machine translation models for generic Basque-Spanish translation, and compare their results with those obtained with rule-based and statistical phrase-based systems (Koehn et al., 2003). Our exploration of NMT variants for this language pair mainly focuses on different sub-word representations, obtained via either linguistically-motivated or frequency-based word segmentations, and on different data exploitation methods. We measure the impact of ordering variations partly via manually-created multiple references and also evaluate the impact of tuning the decoding process in terms of length and coverage along the lines of (Wu et al., 2016).

The paper is organised as follows: Section 2 describes related work in machine translation for Basque and other morphologically-rich languages; Section 3 presents the parallel corpora collected for the Basque-Spanish language pair; Section 4 describes the different translation models used for the experiments presented in Section 5; finally, Section 6 draws conclusion from this work.

2 Related work

Morphologically rich languages, a large denomination which includes synthetic languages where words are formed via productive morphological affixation, have been extensively studied in the machine translation literature. In Statistical Machine Translation (SMT) in particular (Brown et al., 1990), the data sparseness issues created by rich morphology have been addressed with a variety of techniques such as the factor-based translation (Koehn and Hoang, 2007). In Neural Machine Translation, the issues raised by rich morphology are even more acute given the vocabulary limitations for typical encoder-decoder neural networks, and recent work has centred on optimal methods to tackle surface variability and data sparseness in a principled manner.

Several approaches address morphological variation via character-based translation (Ling et al., 2015; Lee et al., 2016; Costa-Jussà and Fonollosa, 2016). A case study along these lines for languages with rich morphology is (Escolano et al., 2017), who implement a character-to-character NMT system augmented with a re-scoring model. They report improvements for Finnish-English translations but not for Turkish-English, although the latter result might be due to lack of sufficient training data.

Other approaches tackle this issue via word segmentation into sub-words. Byte Pair Encoding (BPE) (Sennrich et al., 2016) has become a popular segmentation method where infrequent words are segmented according to character pair frequencies. Alternatives include the use of morphological analysers such as MORFESSOR (Virpioja et al., 2013) or CHIPMUNK (Cotterell et al., 2015). Ding et al. (2016) compare the use of these three segmentation alternatives for Turkish-English, obtaining better results with CHIPMUNK and MORFESSOR than with BPE. In (Ataman et al., 2017), both supervised and unsupervised morphological segmentation are shown to outperform BPE for Turkish to English neural machine translation. Morphological decomposition has also been performed with tools tailored for the task, as is the case in (Sánchez-Cartagena and Toral, 2016), who report improvements using the rule-based morphological segmentation provided by OMORFI (Pirinen, 2015) for English-Finnish translation.

Finally, hybrid techniques have also been applied, as in (Luong and Manning, 2016) who built a character/word hybrid NMT system where translation is performed mostly at the word level and the character component is consulted for rare words. Their results for English to Czech demonstrate that their character models can successfully learn to generate well-formed words for a highly-inflected language. This approach has been notably applied

tion of determiners to the rightmost constituent in the noun phrase.

to English-Finnish by (Östling et al., 2017), who also include BPE segmentation in a system that ranked as the top contribution in the WMT2017 shared task for English-Finnish.

The challenges of machine translation of Basque have been addressed in different frameworks. An example-based data-driven system was thus developed by (Stroppa et al., 2006) and a rule-based approach was used to develop the MATXIN system for Spanish to Basque translation (Mayor et al., 2011); both systems are compared in (Labaka et al., 2007). In (Labaka et al., 2014), a hybrid architecture is presented, combining rule-based and phrase-based statistical machine translation approaches. Their hybrid system resulted in significant improvements over both individual approaches. In the next sections, we provide the first description of a large-scale NMT system for the Basque-Spanish language pair.

3 Corpora

To build representative translation models for the Basque-Spanish language pair, parallel corpora were collected and prepared from three different sources: professional translations in different domains, bilingual web pages, and comparable data in the news domain.

3.1 Parallel data

Parallel data for Basque-Spanish are scarce, the largest repository of such data being the professionally translated administrative texts made available in the Open Data Euskadi repository.[2] Amongst these, the largest collection comes from the translation memories of the *Instituto Vasco de Administración Pública* official translation services, with additional data from the *Diputación Foral de Guipúzcoa*. After filtering duplicate segments and dubious segments, we prepared the ADMIN corpus as our main parallel resource.

Additionally, we included four corpora from different domains. Two of them were created from translation memories, namely the SYNOPSIS corpus, a collection of film synopsis, and the IRRIKA corpus, based on content from the Irrika youth magazine. We also included corpora created via automatic alignment of bilingual documents: EIZIE, a corpus of universal literature, and CONSUMER, a collection of articles from Consumer consumption magazine. The EIZIE alignments were also manually revised to ensure a high quality corpus.

CORPUS	SENTENCES	WORDS	
	ES-EU	ES	EU
ADMIN	963,391	23,413,116	17,802,212
SYNOPSIS	229,464	3,501,711	2,824,807
IRRIKA	5,545	99,319	77,574
EIZIE	94,207	2,506,162	1,775,155
CONSUMER	201,353	3,999,156	3,313,798
TOTAL	1,493,960	33,519,464	25,715,972

Table 1: Parallel corpora statistics (unique segments)

CORPUS	SENTENCES	WORDS	
	ES-EU	ES	EU
CRAWL	1,044,581	19,892,360	15,344,336

Table 2: Crawled corpus statistics (unique segments)

The statistics for all parallel corpora are shown in Table 1.

3.2 Crawled data collection

To complement the high quality parallel data described in the previous section, we created a large parallel corpus from crawled data. We used the PACO2 tool (San Vicente and Manterola, 2012), which performs both crawling and alignment to create parallel resources from web corpora.

For this task, the tool was extended with two major optimisations. First, the crawling component was modified in order to retrieve web content dynamically generated via JavaScript. Secondly, both crawling and alignment processes were parallelised, to speed up the overall process.

Both optimisations contributed to the efficient creation of a parallel corpus from a variety of Basque-Spanish web pages. Corpus statistics, after duplicates removal, are shown in Table 2.

CORPUS	SENTENCES	WORDS	
	ES-EU	ES	EU
EITB	807,222	24,046,414	15,592,995

Table 3: Comparable corpus statistics (unique segments)

3.3 Comparable data collection

To further increase the amount of training data and extend domain coverage, we exploited a large strongly comparable corpus in the news domain, facilitated by the Basque public broadcaster EITB.[3] The original data consists of XML documents that contain news independently created by professional journalists in Basque and Spanish, from

[2] http://opendata.euskadi.eus

[3] http://www.eitb.eus/

CORPUS	SENTENCES	WORDS	
	ES-EU	ES	EU
MERGED	3,345,763	76,998,621	56,391,670
MERGED.LGF	3,086,231	61,529,688	47,976,559

Table 4: Final corpora statistics (unique segments)

which a first parallel corpus was created and shared with the research community (Etchegoyhen et al., 2016).

As additional data was made available since the first version of the corpus, we created a second version that included news from 2013 to 2016. For this version, document alignment was performed with DOCAL (Etchegoyhen and Azpeitia, 2016a), an efficient aligner that provided the best results for this language pair. Sentences were then aligned with STACC (Etchegoyhen and Azpeitia, 2016b), a tool to determine parallel sentences in comparable corpora which has proved highly successful for the task (Azpeitia et al., 2017).

After enforcing one-to-one alignments, the corpus resulted in $1,137,463$ segment pairs, each with an associated alignment probability score. Various corpora could then be extracted by selecting different alignment thresholds to filter low-scoring pairs. After training separate SMT models on each of these three corpora, we selected the corpus with alignment threshold 0.15, as it resulted in the best performance overall. Statistics for this corpus are shown in Table 3.

The EITB corpus was also used to prepare tuning and validation sets, as it covers a wide range of topics that includes politics, culture and sports, among others. Thus, $2,000$ segment pairs were selected as held-off development set, and $1,678$ as test set. The alignments for the test set were manually validated and a new set of references was manually created by professionally translating the Spanish source sentences, to account for word order variability in Basque.[4]

3.4 Data filtering & preparation

A unique parallel corpus (hereafter, MERGED) was built by concatenating the previously described corpora and removing duplicates. All sentences were truecased, with truecasing models trained on the monolingual sides of the bitext, and tokenised with adapted versions of the scripts available in the MOSES toolkit (Koehn et al., 2007).

Neural machine translation systems have been shown to be strongly impacted by noisy data (Belinkov and Bisk, 2017). As our gathered corpora comes from potentially noisy sources, as is the case with crawled and comparable data, we prepared an additional filtered version of the corpus. We based our filtering process on length irregularities between source and target sentences, in terms of number of words, with the aim of identifying those pairs where only a subset of a sentence is translated into the other language, a typical case in comparable corpora.

As a simple approach, we computed the modified z-score on the MERGED corpus to filter out segment pairs identified as outliers in terms of length difference between the source and target segments, where the median and standard deviation were computed on the human quality references of the ADMIN corpus. This process discarded $259,532$ segment pairs, as shown in Table 4, where MERGED.LGF refers to the filtered corpus.

4 Models

In the next subsections, we describe the different NMT models for Basque-Spanish that were built using the corpora described in the previous section, as well as the considered baseline systems.

4.1 Baselines

Two kinds of baseline systems were considered: statistical (SMT) and rule-based (RBMT).

All SMT translation models are phrase-based (Koehn et al., 2003), trained using the Moses toolkit (Koehn et al., 2007) with default hyperparameters and phrases of maximum length 5. Word alignment was performed with the FASTALIGN toolkit (Dyer et al., 2013), and the parameters of the log-linear models were tuned with MERT (Och, 2003). All language models are of order 5, trained with the KENLM toolkit (Heafield, 2011).

As an RBMT baseline translation system, we chose the on-line translation service provided by

[4] In what follows, the manually validated test will be referred to as ALNTEST, the human translations by HTTEST and the multi-reference test set as MULTIREF. Note that all test sets will be made available to the research community on our project web page: `http://modela.eus`.

the Basque Government, which is based on a proprietary rule-based system crafted for this language pair to provide general-domain translation.[5]

4.2 NMT

Unless otherwise specified, all NMT systems follow the attention-based encoder-decoder approach (Bahdanau et al., 2015) and were built with the OPENNMT toolkit (Klein et al., 2017). We use 500 dimensional word embeddings for both source and target embeddings. The encoder and the decoder are 4-layered recurrent neural networks (RNN) with 800 LSTM hidden units and a bidirectional RNN encoder. The maximum vocabulary size was set to $50,000$.

The models were trained using Stochastic Gradient Descent with an initial learning rate of 1 and applying a learning decay of 0.7 after epoch 10 or if no improvement is gained on the loss function after a given epoch over the validation set. A mini-batch of 64 sentences was used for training, with a 0.3 dropout probability applied between recurrent layers and a maximum sentence length set to 50 tokens for both source and target side.

The following subsections describe the neural machine translation variants that were prepared, the first three being based on different word segmentations and the last one on fully character-based translation.

4.2.1 Byte Pair Encoding

Byte Pair Encoding (BPE) is a compression algorithm that was adapted to word segmentation for NMT by (Sennrich et al., 2016). It iteratively replaces the most frequent pair of characters in a sequence with an unused symbol, without considering character pairs that cross word boundaries. BPE allows for the representation of an open vocabulary through a fixed-size vocabulary of variable-length character sequences, having the advantage of producing symbol sequences still interpretable as sub-word units.

For our experiments, we trained joint BPE models on both Basque and Spanish data to improve consistency between source and target segmentation. A set of at most $30,000$ BPE merge operations was learned for each training corpus.

4.2.2 FLATCATV2

FLATCATV2 is a system based on MORFESSOR that was developed to implement a linguistically motivated vocabulary reduction for neural machine translation and was originally proposed for Turkish (Ataman et al., 2017). The segmentation process consists of two steps. MORFESSOR is used first to infer the morphology of the considered language in an unsupervised manner, based on an unlabelled monolingual corpus. The learned morphological segmentations are then fit into a fixed-size vocabulary, which amounted to $45,000$ tokens in our case.

Unlike the joint learning method we selected for BPE segmentation, FLATCATV2 segmentation was learned on the monolingual data separately, since this technique is designed to extract a linguistically-sound segmentation of the text.

4.2.3 Morphological analysis

As a third approach to word representation, we opted for a fine-grained morphological analysis and used the IXA-KAT supervised morphological analyser for Basque (Alegria et al., 1996; Otegi et al., 2016). This analyser relies on a lexicon crafted by linguists which includes most of the Basque morphemes and is used to extract all possible segmentations of a word. The hypotheses with the longest lemma are ultimately selected.

Although this linguistically-motivated approach to segmentation does reduce the vocabulary, vocabulary size is not guaranteed to remain within the range necessary for NMT. We therefore followed the two-step approach used in FLATCATV2 and applied BPE after the linguistic segmentation phase, to segment rare tokens and keep the vocabulary within the selected size.

4.2.4 Character-based translation

As an alternative to NMT architectures based on words or sub-words, character-based models provide the means to remove the segmentation problem altogether. These models are based solely on the characters in the sentence on both the input and the target sides, generating translations one character at the time. As previously discussed, this type of approach is particularly interesting for highly inflected languages such as Basque.

To evaluate this approach for Basque-Spanish translation, we built a character-to-character system following (Lee et al., 2016), whose code was publicly available.[6] The system uses convolutional

[5] http://www.itzultzailea.euskadi.eus/traductor/portalExterno/text.do

[6] https://github.com/nyu-dl/dl4mt-c2c

neural networks to generate window representations of fixed length character sequences, set to 5 in our configuration. These representations reduce the length of the input sequence, while enabling the system to identify segment patterns. A bi-directional recurrent neural network is then used to compute the representation of the complete sentence. Finally, translation is generated character by character, using an attention mechanism on the segments computed at the encoder level.

5 Experiments

In this section we first describe the experimental settings and system variants, then present and discuss the results.

5.1 Settings

To compare the different segmentation approaches, a first set of experiments was designed using only the selected EITB corpus. This allowed for a direct comparison between the approaches while also reducing the computational load of training the different variants. From this set of experiments, we selected the overall best approach to segmentation, taking into account the results obtained in both translation directions.

The second set of experiments compares NMT variants, based on the selected segmentation approach, to the SMT system. We also compare the NMT and SMT results with those obtained with the selected rule-based system on the single and multi-reference test sets.

The NMT approach based on the selected segmentation mechanism was trained on the entire corpus, as was the SMT system. Additionally, we evaluated the same NMT architecture and trained a model on the filtered corpus to assess the impact of noisy data on the final system.

We also evaluated the impact of the decoding optimisations proposed in (Wu et al., 2016), which tune the decoder according to length normalisation over the generated beam sequences and to the coverage of the input sequence according to the attention module. We also tuned the decoder with the end of sentence (EOS) penalty available as hyper-parameter in OPENNMT. Optimal parameters for these three normalisations were evaluated via grid search, resulting in values of 4 for length, 0 for coverage, and 10 for EOS normalisations in ES-EU, and 10, 0 and 10 respectively in EU-ES.

Finally, we performed a small manual evaluation using the Appraise tool (Federmann, 2012). 28 native speakers of Basque were asked to select their preferred translations for 100 sentences, where the translations were generated by the previously described RBMT system and the NMT system trained on the entire corpus.

5.2 Results

Results in terms of automatic metrics were computed with BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). Tables 5 and 6 show the results for the different approaches to segmentation.[7]

The first noticeable result is the consistency of the scores across all test sets, in both directions. For ES-EU, there was no significant difference between the results obtained with BPE and with the unsegmented words, both achieving the best scores overall. In EU-ES, the optimal approach consistently involved applied linguistically-motivated segmentation first, followed by BPE to restrict the vocabulary size. In both directions, FLATCATV2 performed worse than BPE and character-based translation resulted in the lowest scores overall.

Linguistically-motivated segmentation for Basque was only beneficial on the source side, resulting in degraded results when compared to frequency-based segmentation on the target side. This result may be attributed to the stronger need to disambiguate source-side information in NMT architectures, where weak encoding impacts both sentence representation and the attention mechanism. As conditioned language models, NMT decoders seem to have lesser difficulties in generating correct output on the basis of non-linguistic but consistent segmentation units of the type provided by BPE.

From this first set of results, we selected BPE as our segmentation model for the final systems trained on the entire corpus, as it provided the best results when translating into Basque, was a competitive second ranked system in the other translation direction, and required less resources overall to perform segmentation. The comparative results between, RBMT, SMT and NMT are shown in Tables 7 and 8.[8]

[7] In both tables, † indicates statistical significance between the considered system and BPE, for $p < 0.05$. Significance was computed only for the BLEU metric, via bootstrap resampling (Koehn, 2004).

[8] In both tables, † indicates statistical significance between the considered system and NMT, for $p < 0.05$. Results are given on cased and tokenised output translations, after tokenising the output of the RBMT system for a fair comparison.

SEGMENTATION	VOCABULARY	ALNTEST		HTTEST		MULTIREF	
		BLEU	TER	BLEU	TER	BLEU	TER
WORDS	**50,004 / 50,004**	**19.82**	**64.84**	**18.53†**	**61.51**	**28.72**	**55.71**
BPE	**21,765 / 23,741**	**19.51**	**64.65**	**18.00**	**62.20**	**28.40**	**56.00**
FLATCATV2	38,653 / 29,860	18,23†	65,58	17.43†	62.58	27.13†	56.51
FLATCAT (ES) - MORF (EU)	38,653 / 50,004	16.98†	66.66	16.01†	64.09	25.32†	57.99
BPE (ES) - MORF+BPE (EU)	39,197 / 38,827	18.70†	65.31	17.51†	62.64	27.62†	56.36
CHARNMT	304 / 302	17.17†	67.59	16.23†	64.30	25.04†	59.01

Table 5: Evaluation results of the ES-EU systems using different data segmentation on the EITB corpus

SEGMENTATION	VOCABULARY	ALNTEST		HTTEST	
		BLEU	TER	BLEU	TER
WORDS	50,004 / 50,004	26.40	58.82	33.64†	50.08
BPE	23,741 / 21,765	26.61	58.16	35.71	47.67
FLATCATV2	29,860 / 38,653	24.46†	59.88	32.54†	50.43
MORF (EU) - FLATCAT (ES)	50,004 / 38,653	23.90†	60.80	31.06†	51.78
MORF+BPE (EU) - BPE (ES)	**38,827 / 39,197**	**27.86†**	**56.97**	**37.23†**	**46.33**
CHARNMT	304 / 302	24.58†	64.40	31.59†	57.66

Table 6: Evaluation results for EU-ES systems with different data segmentation on the EITB corpus

In Spanish to Basque, when considering all test sets, the best NMT system outperformed SMT, which in turn provided markedly better results than the RBMT system. Interestingly, the SMT system obtained the best BLEU score on the ALNTEST dataset, and was competitive with the basic NMT system for this metric on the MULTIREF test set as well, while being systematically outperformed on the TER metric by all NMT variants. These results might be due to the known BLEU bias in favour of SMT output, along with other biases (Callison-Burch et al., 2006), and the overall results therefore need to be interpreted by considering both metrics in conjunction. Thus, overall NMT performed markedly better, with gains above 4 BLEU points and 5 TER points on the MULTIREF metric. These constitute significant improvements, indicating that NMT responds better to the challenging properties of Basque than alternative approaches.

For Basque to Spanish translation, the comparative results were similar in terms of systems ranking and in terms of larger differences when considering human translations, used as source for this translation direction. As is usually the case, scores were higher when translating into the language with relatively simpler morphology.

Removing noise from the training corpus, via filtering outliers in terms of length differences, had a significant impact on ES-EU, where the MERGED.LGF model outperformed the non-filtered model by close to 3 BLEU points and 2 TER points on the MULTIREF test set. This confirms the importance of a careful preparation of training data for NMT models. For EU-ES, the filtered corpus gave statistically significant improvements as well, although by a lower margin.

Manual examination of the translations produced by the NMT system indicated that lost-in-NMT-translation phenomena, where the system ignores a significant portion of the input sentence in favour of a fluent but incomplete translation, were notable. The MERGED.LGF.OPT version of the system, where output generation is controlled via the previously described normalisation settings, improved on these grounds, both in terms of metrics and after manual examination of a subset of translations where coverage of the source content seemed to improve.

Another interesting aspect in these results is the impact of multiple references on the interpretation of the results. In most cases, taking into account only the initial test set based on alignments, all validated by human experts as proper translations, would have led to different conclusions than those reached when considering both the additional human translations and multiple references. One could have concluded, for instance, that the gains obtained for ES-EU with NMT over SMT were minor, when the differences were much larger overall when considering all references. The need for multiple references in general, and for this language pair in particular, is made even clearer from the results of these experiments.

Finally, Table 9 shows the results from the comparative human evaluation. Overall, users showed a marked preference for the translations produced by the NMT system, selecting RBMT translations in only 15.14% of the cases. Inter-annotator agree-

SYSTEM	ALNTEST		HTTEST		MULTIREF	
	BLEU	TER	BLEU	TER	BLEU	TER
RBMT	9.08†	79.90	14.01†	66.08	17.17†	66.37
SMT	**23.63†**	65.24	17.40†	61.66	30.43	56.50
NMT	20.46	64.52	23.63	55.39	31.27	53.54
NMT.LGF	22.09†	63.36	23.10†	55.10	34.17†	51.73
NMT.LGF.OPT	22.33†	**63.48**	**23.69†**	**54.47**	**34.65†**	**51.42**

Table 7: Final system evaluation results for ES-EU

SYSTEM	ALNTEST		HTTEST	
	BLEU	TER	BLEU	TER
RBMT	16.76†	69.28	25.06†	58.07
SMT	28.09	60.20	32.46†	52.79
NMT	27.68	55.37	39.21	42.58
NMT.LGF	27.99	55.09	39.73†	42.00
NMT.LGF.OPT	**29.02†**	**54.36**	**40.56†**	**41.26**

Table 8: Final system evaluation results for EU-ES

NMT>RBMT	NMT=RBMT	RBMT>NMT	SKIPPED
67.64%	15.39%	15.14%	1.82%

Table 9: Human evaluation results for ES-EU

ment measures showed fair agreement, with 0.306 and 0.309 for the Krippendorf's Alpha and Fleiss' Kappa metrics, respectively. Although admittedly based on a small sample, these results confirmed the impressions from users of the NMT system, who characterised it as a significant step forward in machine translation of Basque.

6 Conclusions

We presented the first results in neural machine translation for Basque, a synthetic language with an extended case system, complex verbal morphology and relatively free word order. The characteristics of the language made it an interesting test case for NMT and we showed that significant gains could be obtained with a neural network approach, when compared to both rule-based and statistical machine translation systems.

Several variants were prepared in terms of both corpora and models, to determine the optimal configurations for generic machine translation in Basque-Spanish. The impact of noisy datasets when training NMT systems was confirmed in our experiments, and we showed the improvements achievable with a simple filtering of length difference outliers.

Also coming from our results were the gains resulting from fine-grained morphological analysis on the source side, although byte pair encoding was shown to be a robust method overall for this language pair. The presented results were computed on different complementary test set, providing a strong confirmation of the importance of multiple references in general, and for the evaluation of Basque translation in particular.

Neural machine translation has been successfully applied to a large range of languages and scenarios, with recent work centred on languages with rich morphology. This work contributes to this line of research, demonstrating the significant improvements obtained with NMT on a language which features a wide range of properties that represent a challenge for past and current approaches to machine translation.

Acknowledgements: This work was supported by the Department of Economic Development and Competitiveness of the Basque Government via the MODELA project. We would like to thank the three anonymous EAMT reviewers for their helpful comments and suggestions, as well as all participants of the manual evaluation.

References

Alegria, Iñaki, Xabier Artola, Kepa Sarasola, and Miriam Urkia. 1996. Automatic morphological analysis of Basque. *Literary and Linguistic Computing*, 11(4):193–203.

Ataman, Duygu, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108:331–342.

Azpeitia, Andoni, Thierry Etchegoyhen, and Eva Martinez Garcia. 2017. Weighted Set-Theoretic Alignment of Comparable Sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Belinkov, Yonatan and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.

Bojar, Ondřej et al. 2016. Findings of the 2016 conference on machine translation (WMT2016). In *Proceedings of the First Conference on Machine Translation*, WMT2016, pages 131–198, Berlin, Germany.

Bojar, Ondřej et al. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, WMT2017, pages 169–214, Copenhagen, Denmark.

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

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

Costa-Jussà, Marta R. and José A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 357–361, Berlin,Germany.

Cotterell, Ryan, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 164–174, Beijing, China.

Crego, Josep, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

De Rijk, Rudolf PG and Armand De Coene. 2008. *Standard Basque: A progressive grammar*. MIT Press Cambridge, MA.

Ding, Shuoyang, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post. 2016. The JHU machine translation systems for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, WMT2016, pages 272–280, Berlin, Germany.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 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*.

Escolano, Carlos, Marta R. Costa-jussà, and José A. R. Fonollosa. 2017. The TALP-UPC Neural Machine Translation System for German/Finnish-English Using the Inverse Direction Model in Rescoring. In *Proceedings of the Second Conference on Machine Translation*, WMT2017, pages 283–287, Copenhagen, Denmark.

Etchegoyhen, Thierry and Andoni Azpeitia. 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, Thierry and Andoni Azpeitia. 2016b. Set-Theoretic Alignment for Comparable Corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1: Long Papers, pages 2009–2018, Berlin, Germany.

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

Federmann, Christian. 2012. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Gehring, Jonas, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Heafield, Kenneth. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT 2011, pages 187–197, Edinburgh, Scotland.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1701.02810*.

Koehn, Philipp and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 joint conference on Empirical Methods in Natural Language processing and Computational Natural Language Learning (EMNLP, CoNLL)*, pages 868–876, Prague, Czech Republic.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics, (HLT-NAACL)*, pages 48–54, Edmonton, Canada.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Prague, Czech Republic.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.

Labaka, Gorka, Nicolas Stroppa, Andy Way, and Kepa Sarasola. 2007. Comparing Rule-Based and Data-Driven Approaches to Spanish-to-Basque Machine Translation. In *Proceedings of MT-Summit XI*, pages 297–304.

Labaka, Gorka, Cristina España-Bonet, Lluís Màrquez, and Kepa Sarasola. 2014. A hybrid machine translation architecture guided by syntax. *Machine Translation*, 28:91–125.

Lee, Jason, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Ling, Wang, Isabel Trancoso, Chris Dyer, and Alan Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Luong, Minh-Thang and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. *arXiv preprint arXiv:1604.00788*.

Mayor, Aingeru, Iñaki Alegria, Arantza Díaz de Ilarraza, Gorka Labaka, Mikel Lersundi, and Kepa Sarasola. 2011. Matxin, an open-source rule-based machine translation system for basque. *Machine Translation*, 25:53–82.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, ACL '03, pages 160–167, Sapporo, Japan.

Östling, Robert, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The Helsinki Neural Machine Translation System. In *Proceedings of the Second Conference on Machine Translation*, WMT2017, pages 338–347, Copenhagen, Denmark.

Otegi, Arantxa, Nerea Ezeiza, Iakes Goenaga, and Gorka Labaka, 2016. *A Modular Chain of NLP Tools for Basque*, pages 93–100. Springer International Publishing.

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

Pirinen, Tommi A. 2015. Development and use of computational morphology of Finnish in the open source and open science era: Notes on experiences with OMorFi development. *SKY Journal of Linguistics*, 28:381–393.

San Vicente, Inaki and Iker Manterola. 2012. Paco2: A fully automated tool for gathering parallel corpora from the web. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC2012, pages 1–6, Istanbul, Turkey.

Sánchez-Cartagena, Víctor M. and Antonio Toral. 2016. Abu-MaTran at WMT 2016 Translation Task: Deep Learning, Morphological Segmentation and Tuning on Character Sequences. In *Proceedings of the 1st Conference on Machine Translation*, WMT2016, Berlin, Germany.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1715–1725, Berlin,Germany.

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

Stroppa, Nicolas, Decan Groves, Andy Way, and Kepa Sarasola. 2006. Example-based machine translation of the basque language. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 232–241, Cambridge, MA USA.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, NIPS, pages 3104–3112, Montreal, Canada.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, NIPS, pages 6000–6010, Montreal,Canada.

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline. *Aalto University publication series SCIENCE + TECHNOLOGY; 25/2013*.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.