EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

PhD thesis summary

Corpus compilation and development of a machine translation system for translating clinical reports between Basque and Spanish

Xabier Soto Garcia

2021

EUSKAL HERRIKO UNIBERTSITATEA
University of the Basque Country

Corpus compilation and development of a machine translation system for translating clinical reports between Basque and Spanish

This is a shortened version of the Basque dissertation entitled *Txosten klinikoak euskararen eta gaztelerarren artean itzultzen laguntzeko corpusaren bilketa eta itzultzaile automatikoaren garapena*, written by Xabier Soto Garcia under the supervision of Dr. Maite Oronoz Anchordoqui and Dr. Gorka Labaka Intxauspe. It also includes the papers published by the candidate on the research presented here.

November 2021

Acknowledgments

The Spanish Ministry of Economy and Competitiveness, who awarded me a predoctoral fellowship (BES-2017-081045) to conduct research within the PROSAMED project (TIN2016-77820-C3-1-R).

Abstract

This dissertation summarizes the work done for developing a Machine Translation (MT) system for translating clinical reports between Basque and Spanish. With the aim of promoting the use of Basque language when writing clinical reports, we prioritize the Basque-to-Spanish (eu-es) translation direction.

Our approach is data-centric, focusing on the compilation of diverse corpora that can be useful for translating clinical reports between Basque and Spanish. Given that we have access to many health records in Spanish, we incorporate them into our systems through back-translation when translating from Basque to Spanish.

One of the main characteristics of the clinical domain is its rich terminology, so when compiling the corpora we have payed particular attention to the clinical terminologies available in Basque and Spanish. Then, we have used these terminologies both for training our systems and for performing semi-automatic error analyses in some of our systems.

While gathering data, we tried different MT systems, architectures and techniques. For developing our final systems, we made use of Neural Machine Translation (NMT), but for some of our experiments we also tried Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) for back-translation. Regarding NMT architectures, we trained Transformer models and Recurrent Neural Networks (RNNs) of different size. We also carried out different preprocessing approaches, including the application of different word segmentation methods.

Apart from evaluating the MT quality of the developed systems, we also measured the lexical diversity of the corpora created by some of the back-translation systems, trying to link the lexical diversity of the source side of the training corpus and the MT quality of the systems trained with this corpus. Within this analysis, we also measure the gender bias of our bilingual clinical domain corpus, counting the number of appearances of the terms 'nurse' and 'doctor' in their masculine and feminine forms in Spanish.

For measuring the environmental impact of our work, we calculate the power consumed when training some of our systems, and estimate the corresponding CO_2 emissions.

Finally, for studying the generalizability of our proposed methods, we repeat some of our experiments in other language pairs with publicly available data.

Contents

1 Introduction

The MT systems developed in this PhD aim to be implemented in Osakidetza (the public health service in the Basque Autonomous Community), so they are designed to translate between Basque and Spanish. Nowadays, given that not all of the healthcare workers in Osakidetza understand Basque, and the health records are stored in a centralized way, most of the healthcare workers write their clinical reports in Spanish. Thus, with the aim of helping Basque speaking healthcare workers to write their health records in Basque, we focus on the development of an MT system for translating clinical reports from Basque to Spanish.

This work is part of the Itzulbide project, which has 3 objectives: 1) the compilation of bilingual clinical reports; 2) the development of MT systems to translate clinical reports between Basque and Spanish; and 3) the human evaluation of the developed systems by the same volunteers who compiled the bilingual corpus. This project was presented by Osakidetza and won by the Ixa group, where this work is carried out. The research area of Ixa group is Natural Language Processing (NLP), being MT one of its sub-areas.

The main part of this introduction will mention the different MT systems and NMT architectures used in this work, along with some of the preprocessing techniques and back-translation approaches tried during this PhD. Later, we will briefly describe two areas explored in this work, namely data selection and lexical diversity.

1.1 Machine Translation

Machine Translation is defined as the task of automatically translating a text in a given language to a text with the same meaning in another language. Historically, diverse MT approaches have been explored.

Initially, the focus was on linguistics, trying to define rules for translating the text from the source to the target language, considering the characteristics of each language, and making use of bilingual dictionaries. This is known as Rule-Based Machine Translation (RBMT), and the specific system we use in this work for translating between Basque and Spanish is *Matxin* (Mayor, 2007).

Further on, with the increase in the amount of digitalized texts, corpus based systems were developed and became the state-of-the-art systems in the 1990s. Firstly, Statistical Machine Translation (SMT) systems were proposed (Koehn et al., 2003), which are based on counting the appearances of each word/phrase in each language of the bilingual corpus, and automatically inferring which words/phrases convey the same meaning. These systems were also adapted for the translation between Basque and Spanish, by the system known as *EUSMT* (Labaka, 2010).

Later on, with the access to more computational capabilities in the 2010s,

Neural Machine Translation (NMT) displaced SMT as the state-of-the-art technique (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015). NMT systems use an encoder-decoder approach, with one network learning to encode the meaning of the source sentence, and another one designed to decode the meaning of the source sentence into the target language, using embeddings for representing words and sentences. These systems have also been tried for translating between Basque and Spanish, being *MODELA* (Etchegoyhen et al., 2018) the first attempt in this respect.

Within NMT, different architectures have been proposed. The first systems were based on RNNs, but nowadays most of the NMT systems use a Transformer (Vaswani et al., 2017) architecture. RNNs sequentially read the source sentence, changing its representation after reading every word. Thus, the encoding of the source sentence is highly dependent on the first word, making it hard to translate long sentences or sentences in which there is any linguistic relation between distant words. Attention mechanism (Bahdanau et al., 2015) was proposed to counter this problem, allowing the systems to predict which are the words in the source sentence most relevant when generating each word in the target sentence. Transformer is purely based on the attention mechanism, learning the relations between each word in a sentence with the rest of the words in that sentence, regardless of their position. This allows the systems to learn more complex relations, whatever the distance between the words that form this relation.

Regardless of the architecture, one important aspect for improving the results of NMT is preprocessing. Apart from the usual tokenization and Truecasing, word segmentation is needed for enabling the systems to work with a limited vocabulary. With that purpose, a well-known data compression method, namely Byte Pair Encoding (BPE), was adapted for working with written text (Sennrich et al., 2015). By this method, the corpus is initially divided into characters, and then the most frequent character group is merged until a given number of iterations are completed. This technique is specially useful for morphologically rich languages as Basque. Later, it was proposed to apply regularization to this word segmentation approach, giving birth to BPE-dropout (Provilkov et al., 2020). This regularization is also part of *sentence-piece*[1], a method equivalent to BPE.

Another technique that helped NMT to become state-of-the-art is back-translation (Sennrich et al., 2016). This technique is based on a simple idea: for a given language pair, automatically translate a corpus in the target language into the source language, increasing the number of sentences available for training a system for that language pair. This allows leveraging monolingual corpora in the target language, giving the possibility of adapting systems to a given domain when there are no texts of that domain in the source language.

[1] `https://github.com/google/sentencepiece`

Typically, beam search (Tillmann and Ney, 2003) is used for decoding the target sentence, by considering only the most probable outputs until the whole sentence is generated. However, sampling was shown to obtain better results when back-translating monolingual corpora (Edunov et al., 2018); and later two different approaches were proposed for further improving the back-translation results: 1) restricting the set of words that can be chosen when sampling to the most probable ones or the ones above a certain probability threshold (Graça et al., 2019); and 2) tagging the corpus created through back-translation for helping the system to distinguish the synthetic corpora from the bilingual one (Caswell et al., 2019).

Overall, the state-of-the-art in MT is defined at the Conference on Machine Translation (WMT), and among the diverse domains studied, translation of news is the one that attracts more participants. Regarding language pairs, English-to-German and English-to-French are usually taken as the main reference.[2]

For evaluating the MT systems, human evaluation is preferred whenever possible, commonly measuring the fluency and accuracy of the generated translations. For doing that, evaluators are usually asked to rate each translation on a 1-5 scale, known as *Likert* scale. However, considering the high cost of human evaluation, automatic metrics are defined for boosting the development of MT systems. In this work, we use BLEU (Papineni et al., 2002), TER (Snover et al., 2006), METEOR (Banerjee and Lavie, 2005) and chrF (Popović, 2015) as metrics.

1.2 Lexical Diversity

MT systems tend to produce texts with lower lexical diversity than the original source sentences (Vanmassenhove et al., 2019). This is partly owing to the use of decoding algorithms like beam search that only consider the most probable output words at each decoding step; and indirectly, also because of the use of precision-based metrics like BLEU for choosing the best system to be implemented.

Therefore, when comparing the performance of different back-translation approaches, apart from measuring the MT quality of the systems trained with that back-translated corpora, we also calculate the lexical diversity of the corpora created through each of the methods under study. For doing this, we use the metrics TTR (Type-Token Ratio) (Templin, 1975), Yules' I (Yule, 1944) and MTLD (Measure of Textual, Lexical Diversity) (McCarthy, 2005).

[2] `https://github.com/sebastianruder/NLP-progress/blob/master/english/machine_translation.md`

1.3 Data Selection

When developing an MT system, the usual strategy is to collect a corpus from the desired language pair and domain that contains as many sentences as possible. However, there is a research area called data selection that aims to reduce the training corpus while keeping or even improving the MT quality. For this purpose, the sentences from the available corpus most similar to the desired task are selected, reducing the training time of the systems and their consumed power.

Among the different data selection approaches, we choose Feature Decay Algorithms (FDA) (Biçici and Yuret, 2015; Poncelas et al., 2018a) for being the most adequate for MT (Silva et al., 2018). This method is based on selecting the sentences most similar to the sentences from the development set. To this end, in each iteration the sentence with higher n-gram overlap with the sentences from the development set is selected. At the same time, for making the selected corpus more diverse, when selecting a new sentence the n-grams that have been selected most times are penalized.

2 Resources

This section briefly describes the resources employed during this PhD. First, we define the different corpora used for training and evaluating our systems. Finally, we mention the different MT systems we have used, along with the techniques applied for preprocessing.

2.1 Corpora

For training the MT systems to translate between Basque and Spanish we use four types of data: 1) out-of-domain bilingual corpora, 2) bilingual clinical terminology, 3) bilingual clinical corpora, and 4) monolingual clinical corpora in Spanish. The first 3 types of corpora are used to train the Spanish-to-Basque systems, which are used to back-translate the monolingual clinical corpora, and all together are used for training the Basque-to-Spanish systems. The bilingual clinical corpora are used for fine-tuning and evaluating the systems in each direction.

Regarding the out-of-domain bilingual corpora, we have looked for a balance between gathering as much data as possible and guaranteeing a minimum quality. With that in mind, we have selected corpora that, whether have been previously used for training MT systems between Basque and Spanish, or have been translated by professional translators and have similar characteristics to the clinical texts which are the focus of this PhD.

Table 1 lists the diverse out-of-domain bilingual corpora used in this work, specifying their domain and number of sentences. Note that not all the

corpora were available from the beginning of this PhD, so different out-of-domain corpora have been used in each publication. For the final systems, the first 7 corpora listed in table 1 were included after applying a language identification tool[3]; the repeated sentences coming from other corpora were deleted, and the sentences longer than 100 tokens were removed by applying a corpus cleaning tool[4]. After this, around 5M out-of-domain sentences were used for training our final systems.

Corpus	**Domain**	**Number of sentences**
EiTB (2016) (Etchegoyhen et al., 2016)	News	0.56M (x3)
HAEE	Administrative	0.9M
Consumer	Consumerism	268,112
Irrika	Science divulgation	5,570
EIZIE	Translation memories	94,552
Pelikulen_sinopsiak	Film synopses	237,883
PacoWebCorpus2012 (San Vicente and Manterola, 2012)	Web-crawling	659,395
HAC (Sarasola et al., 2015)	Literature	566,738
Osakidetza_professionals	Health/administrative	22,051
EiTB (2020) (Etchegoyhen and Gete, 2020)	News	637,182

Table 1 – Out-of-domain bilingual corpora, indicating their domain and number of sentences.

Most of the clinical terminology used in this PhD comes from SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) (IHTSDO, 2014), considered the most comprehensive clinical terminology collection in the world. This terminology was automatically translated into Basque (Perez-de Viñaspre, 2017), combining the use of dictionaries, transliteration of neoclassic terms, generation of nested terms based on predefined rules and adaptation of a RBMT system to the medical domain.

Another terminology source used in this work is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10), manually translated into Basque and made available in the WMT Biomedical shared task (Bawden et al., 2020).[5]

Finally, a few terms related to COVID-19 have been compiled, one coming from an interim release of SNOMED CT[6], translated into Basque by a

[3] https://github.com/saffsd/langid.py

[4] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

[5] https://drive.google.com/drive/folders/1gUQHoutvYIXGGPVTBbBF3qlHHhX9qbrO

[6] http://www.snomed.org/news-and-events/articles/

professional translator from Osakidetza; and the other coming from a compilation made by Elhuyar foundation and published in their website.[7]

Table 2 sums up the statistics of the bilingual clinical terminology used in this PhD. Note that 2 versions of the SNOMED CT and ICD-10 terminologies were used in different publications, and the numbers shown here correspond to the latest versions.

Terminology	**Terms**	**Tokens (eu)**	**Tokens (es)**
SNOMED CT	896,898	3,074,750	5,309,227
ICD-10	29,670	245,150	188,233
SNOMED CT / COVID-19	84	579	729
Elhuyar / COVID-19	126	263	243

Table 2 – Bilingual clinical terminology, showing the number of terms and tokens in each language.

Three different bilingual clinical domain corpora were used in this work. The first and most important is the one compiled as part of the Itzulbide project, from where the sentences to evaluate the final systems were extracted, leaving the rest for fine-tuning. Also for fine-tuning, a smaller corpus included in the E3C project (Magnini et al., 2020) was added, formed by clinical cases compiled in Basurto University Hospital. This corpus is available on the web[8]. Finally, before having access to the bilingual clinical domain corpus compiled in the Itzulbide project, sentences extracted from health record templates in Basque (Joanes Etxeberri Saria V. Edizioa, 2014), written in Donostia University Hospital and manually translated into Spanish by a doctor from Osakidetza, were used for evaluation.

Table 3 shows the statistics of the bilingual clinical domain corpora used in this work. During this PhD we used 2 versions of the Itzulbide corpus, corresponding the numbers included here to the latest version.

Corpus	**Sentences**	**Tokens (eu)**	**Tokens (es)**
Itzulbide	30,805	353,986	392,607
Basurto University Hospital	541	5,254	5,185
Donostia University Hospital	2,076	19,938	19,022

Table 3 – Bilingual clinical domain corpora, showing the number of sentences and tokens in each language.

Finally, diverse monolingual clinical corpora in Spanish were used in this work, mostly for back-translation. Some of the corpora come directly from Galdakao-Usansolo Hospital and Basurto University Hospital, in the first

`march-2020-interim-snomedct-release%2DCOVID-19`

[7]This page is no longer available.

[8]`https://github.com/hltfbk/E3C-Corpus`

case formed by discharge reports, and in the second containing also progress reports. The rest of the clinical domain corpus in Spanish was also part of the Itzulbide project, and given its big size, 5M sentences coming from this source were selected and added to the previous corpora. This way, the number of sentences used for back-translation in the final systems was similar to the double of the number of bilingual sentence pairs, ensuring that back-translation gives the expected improvement (Poncelas et al., 2018b).

Table 4 presents the statistics and document type of the monolingual clinical domain corpora used in this work. The numbers shown here correspond to the original corpora, before removing empty sentences, repeated sentences, or sentences containing only codes and dates. Note also that, from the different corpora coming from Itzulbide, only the progress reports and the ones from trauma specialty were used after applying data selection.

Corpus	**Document type**	**Sentences**	**Tokens (es)**
Galdakao-Usansolo Hospital	Discharge reports	4.363.627	47.417.680
Basurto University Hospital	Discharge reports	2.713.424	17.144.473
Basurto University Hospital	Progress reports	4.811.294	29.047.905
Itzulbide (es)	Progress reports	49,069,600	270.753.011
Itzulbide (es)	Trauma specialty	2,412,202	12,521,055
Itzulbide (es)	Hospitalization	6,550,241	43,761,415
Itzulbide (es)	Emergencies	18,576,314	97,488,753

Table 4 – Monolingual clinical domain corpora, showing the document type, along with the number of sentences and tokens.

2.2 Systems

In this section we specify the different RBMT, SMT and NMT systems used in this PhD, along with the scripts employed for preprocessing.

For RBMT, when translating between Basque and Spanish we use *Matxin-Med*, the adaptation of *Matxin* (Mayor, 2007) to the medical domain by the inclusion of clinical dictionaries. For translating between German and English, we use *Apertium* (Forcada et al., 2011), an open source code for RBMT.

For SMT, we use *Moses* (Koehn et al., 2007) for all language pairs with its default parameters. For word alignment we use *MGIZA* (Och and Ney, 2003), as reordering model we employ an "msd-bidirectional-fe" lexicalised model, and for the target language model we train a 5-gram sized *KenLM* (Heafield, 2011) model. The weights for the different components were tuned to optimize BLEU using Minimum Error Rate Training (MERT) (Och, 2003) with an n-best list of size 100.

For NMT we used different systems throughout this PhD. In the beginning, we used *Nematus* (Sennrich et al., 2017) for training the RNNs of

different size. For the shallow RNN we made our first trials with the hyperparameters employed in *MODELA* (Etchegoyhen et al., 2018); while for the deep RNN (Barone et al., 2017) we used the configuration that gave the best results.[9]

Once it was announced that *Nematus* would not be updated anymore, we started to use *OpenNMT* (Klein et al., 2017), specifically in its PyTorch implementation. Most of the times we used the Transformer architecture with the recommended hyperparameters,[10] and for some experiments we used LSTM (Hochreiter and Schmidhuber, 1997) architecture, with 4 layers, 512 neurons per layer, dropout with 0.2 probability, and a batch-size of 128.

Finally, for training the final systems we used *Fairseq* (Ott et al., 2019), always with the Transformer architecture, and the hyperparameters recommended in the end of this blog post.[11]

Regarding preprocessing, for the corpus based systems we applied the tokenization[12] and *Truecaser*[13] tools available in *Moses*. The Truecase model was learnt on the first 7 out-of-domain corpora presented in Table 1. When using NMT, we applied BPE in its original implementation.[14] In the latest experiments, we saved distinct dictionaries for each language, as recommended by the authors. For some experiments, we also tried BPE-dropout (Provilkov et al., 2020) with 0.1 probability. In all the experiments done between Basque and Spanish, BPE was applied for 90,000 iterations; for translating between English and Spanish this value was 32,000; and for German/English we used 89,500 iterations in the first experiment and 40,000 in the second.

Lastly, for some of the latest experiments we also used the corpus cleaning tool[15] available in *Moses*; whether for limiting the number of tokens from the training sentences to 100, or for truncating the sentences to be back-translated to 1,000 tokens, avoiding memory errors.

[9] https://github.com/Avmb/deep-nmt-architectures/blob/master/configs/bideep-bideep-rGRU-large/config.sh

[10] http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model

[11] http://cslab.org/blog/fairseq-basics

[12] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

[13] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl

[14] https://github.com/rsennrich/subword-nmt

[15] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

3 General outline of the dissertation

This is a shortened version of the Basque-written PhD dissertation entitled *Txosten klinikoak euskararen eta gazteleraren artean itzultzen laguntzeko corpusaren bilketa eta itzultzaile automatikoaren garapena*. The work was done within the Ixa research group, working in the fields of Natural Language Processing and Computational Linguistics. The PhD is focused on Machine Translation, and it is also situated in the area of medical domain NLP.

The Basque dissertation is formed by seven chapters, the contents of which are mainly included in the six publications presented in the Appendix. However, for a proper understanding of the work carried out in this PhD, it is recommended to read this summarized dissertation first. To comprehend the advances done in this PhD, the publications in the Appendix are presented in chronological order. It has to be noted that the bilingual clinical domain corpus compiled in the Itzulbide project was not available from the beginning of this PhD, so the conclusions presented in the first publications using health record templates for evaluation should be taken with caution.

The six publications presented in the Appendix are briefly described below. Each publication is assigned a code, that connects with the publication details shown in the first page of the Appendix (page 23).

[P1] Preliminary work testing the first approaches for translating clinical reports between Basque and Spanish. The work presents the results of two set of experiments: 1) a hyperparameter optimization performed with out-of-domain bilingual corpora; and 2) a study of the effect of adding clinical terminologies and a monolingual corpus on the desired task. The clinical terminologies were added both directly and inside artificially created sentences, while the monolingual corpus was added through back-translation and copying. The systems used in this work were one-layered bidirectional RNNs, and a human evaluation of the best performing Basque-to-Spanish system was performed.

[P2] A comparison of different architectures and systems for translating clinical reports between Basque and Spanish. With the appearance of the Transformer architecture, we tested its performance and compared it with the previous RNN and a deeper RNN. This work also includes the use of different MT systems for back-translation, considering the use of RBMT and SMT along with the previous RNN and Transformer NMT architectures.

[P3] A study applying data selection techniques to corpora back-translated by several systems. Different NMT architectures, as well as RBMT and SMT systems, were used for back-translation; and diverse data selection approaches were taken for filtering the corpora. Furthermore, lexical diversity of the corpora produced by each system was measured;

and the data selection results were rescored considering the lexical diversity scores of the back-translated corpora and the MT metrics of the back-translation systems. For this work, we repeated the experiments done between Basque and Spanish with publicly available data for translating between German and English.

[**P4**] Description of our participation in the WMT 2020 Biomedical shared task. That year, English-to-Basque was considered as a language pair for the first time, both for translating biomedical abstracts and clinical terminologies. Having designed a cascade approach, using Spanish as pivot language, we developed an English-to-Spanish system and use the previously compiled corpora for designing a Spanish-to-Basque system. Given that we had to collect English/Spanish corpora, we also created a Spanish-to-English system, and submitted our systems for both directions including these languages. For the first time in this PhD, we measured the power consumed by the GPUs (Graphics Processing Units) used for training the systems, and estimate the corresponding CO_2 emissions.

[**P5**] Book chapter describing the Itzulbide project and presenting the first results using the bilingual corpus derived from it. In this work, we tried several word segmentation approaches, we tested the performance of our system in a held-out specialty, and tried tagging the bilingual clinical domain corpus according to the specialty of each sentence. As a complement to the reported MT scores, we also performed an error analysis, focusing on the correct translation of clinical terminology.

[**P6**] A paper trying diverse approaches for back-translation, result of combining tagging with different decoding algorithms. As done previously, we performed our experiments both for Basque-to-Spanish and German-to-English language pairs. Apart from measuring the MT quality of the diverse back-translation approaches, we also compared the lexical diversity of the corpus created by each system. Related to lexical diversity, we measured the gender bias of the Itzulbide bilingual corpus, by counting the appearances of the terms 'nurse' and 'doctor' in their masculine and feminine forms in Spanish. Finally, as done before, we estimated the carbon footprint derived from our experiments.

4 Objectives, conclusions, contributions and future work

The main objective of this PhD is to develop an MT system for translating clinical reports from Basque to Spanish. With this aim, several objectives were set and studied through the work mentioned in the previous section. These objectives are listed below (Section 4.1). Then, the major conclusions derived from this research are summarized (Section 4.2), followed by the main contributions (Section 4.3). Finally, possible future work is presented (Section 4.4).

4.1 Objectives

The six objectives set in this PhD are listed below, along with a brief explanation.

[O1] To develop an MT system for translating clinical reports from Basque to Spanish.

Being a priority to translate in this direction, and having access to many clinical reports in Spanish, we have the objective of leveraging this monolingual corpora. For that purpose, we translate them automatically into Basque and use them for training the Basque-to-Spanish systems, making use of the back-translation technique. Thus, another intermediate objective is to develop an MT system for translating clinical reports from Spanish to Basque.

[O2] To compile bilingual and monolingual corpora that can be useful for translating clinical reports.

Most of the systems we use in this PhD are corpus based, so we need to compile the necessary bilingual and monolingual corpora for training them. Therefore, another important objective is to compile in-domain bilingual and monolingual corpora, along with out-of-domain bilingual corpora that can be useful for translating clinical reports between Basque and Spanish. Given the importance of translating correctly the clinical terminology, we also set the objective of compiling clinical terminology available in Basque and Spanish, with the aim of testing different ways of integrating this terminology into our systems.

[O3] To compare different MT systems, architectures and techniques.

Being our approach data-centric, we aim to compare the performance of different MT systems, architectures and techniques in our task of translating clinical reports between Basque and Spanish. Among others, we

consider RBMT, SMT and NMT systems for back-translation, different NMT architectures, and diverse word segmentation approaches.

[O4] To analyse the lexical diversity of the corpus created through back-translation.

Apart from measuring the MT performance of our systems, we have the aim of measuring the lexical diversity of the corpus created through different back-translation approaches. Through this analysis, we want to study to what extent the lexical diversity measured on the source side of the training corpus affects the final performance of the NMT systems. Related to this, another objective of the PhD is to quantify the gender bias of the bilingual clinical domain corpus compiled for this work. With this aim, we count the appearances of the terms 'nurse' and 'doctor' in their masculine and feminine forms in Spanish.

[O5] To test our approaches in other language pairs.

With the aim of situating our proposals in an international level, we repeat some of the experiments made in this work between Basque and Spanish in other language pairs. With that objective in mind, we choose the biomedical domain for being similar to the clinical one and having public data, testing the generalizability of our methods in three language pairs: German-to-English, English-to-Spanish, and Spanish-to-English.

[O6] To measure the carbon footprint of our experiments.

Finally, we want to measure the possible effects of training our NMT systems in the environment. With this aim, we calculate the power consumed by our GPUs during training and estimate the CO_2 emissions derived from some of our experiments.

4.2 Conclusions

In the following, we sum up the main conclusions derived from the experiments made through this work, focusing on the ones related to our task of translating clinical reports between Basque and Spanish.

[C1] Without bilingual in-domain corpus, adding clinical terminology directly to the training corpus is helpful for translating clinical notes.

However, integrating this clinical terms into artificially created sentences was not helpful. Furthermore, in the experiments made between English and Spanish for the WMT 2020 Biomedical shared task, adding the clinical terminology directly to the training corpus did not prove to

be useful, reducing the average sentence length of the produced translations. Thus, this conclusion should be tested before applying it to another language pair.

[C2] Transformer is the best architecture for translating clinical reports between Basque and Spanish, also for back-translation.

This conclusion is derived from the experiments we made trying different architectures for our task, including also the use of different NMT architectures, as well as RBMT and SMT systems, for back-translation.

[C3] Without bilingual in-domain corpus, RBMT can be useful for back-translation, improving the results of SMT.

Related to the previous conclusion, even if Transformer obtained the best results in both translation directions, we observed that, once the back-translated data was added to the training corpus, the system using RBMT for back-translation obtained good results too, improving the performance of SMT.

[C4] The bilingual clinical domain corpus compiled in the Itzulbide project is very helpful for translating clinical reports between Basque and Spanish.

Being one of the objectives of this PhD to compile proper corpora for our task, the bilingual clinical domain corpus compiled in the Itzulbide project was the one that improved the results most significantly.

[C5] BPE-dropout applied to both sides of the training corpus is the best word segmentation approach for translating clinical reports between Basque and Spanish.

However, it has to be considered that BPE-dropout needs more time than BPE; on the one hand, for requiring more training epochs, and on the other hand, for the need to be applied to the training corpus once per every epoch. Therefore, we recommend to apply regular BPE to compare the performance of different systems, and use BPE-dropout only for the final systems to be evaluated.

4.3 Contributions

Apart from the conclusions mentioned in the previous section that were useful for our task of translating clinical reports between Basque and Spanish, during this PhD we made some contributions to the MT community:

[C'1] We combined the output of 4 different back-translation systems, improving the results obtained using only one system in the Basque-to-Spanish direction.

Specifically, we combined the output of different NMT architectures, as well as RBMT and SMT systems. This was an advance from previous work, that only combined SMT and NMT systems (Poncelas et al., 2019).

[C'2] For the first time, we applied data selection techniques to the corpus created through back-translation, improving the results obtained with a 4 times bigger corpus in one of the language pairs according to all data selection approaches.

By doing so, we proposed a new way to apply data selection in NMT systems, giving another possibility to reduce the training corpus, and thus, the training time.

[C'3] We proposed a method to rescore the outputs of data selection based on the MT metrics and lexical diversity scores of the back-translation systems.

This way, we opened the possibility of considering the lexical diversity of the source side of the training corpus for improving the results of NMT systems.

[C'4] We combined two back-translation techniques, creating a new one named *tagged restricted sampling*, that obtains similar results to *unrestricted sampling*, considered the state-of-the-art decoding method for back-translation.

Even if the results obtained with *tagged restricted sampling* and *unrestricted sampling* are very similar, we have to point out that our proposal reduces the time for back-translating the monolingual corpora, which can be useful in high resource settings.

[C'5] We measured the gender bias of the bilingual clinical domain corpus, counting the number of appearances of the terms 'nurse' and 'doctor' in their masculine and feminine forms in Spanish. On the other hand, we calculated the power consumed by the GPUs used for training some of our systems, accordingly estimating the CO_2 emissions produced when training our MT systems.

This way, we analized the impact that our systems could have in the society, measuring the gender bias of our corpora and estimating the environmental impact of training some of our systems, providing some reference numbers that can be considered by other MT researchers.

4.4 Future work

The work done in this PhD can be extended in several ways:

- Get feedback from the users of the MT systems and, if necessary, update them.
- As a continuation of the gender bias study done in the bilingual clinical domain corpus, collaborate with Osakidetza for seeking solutions to prevent our MT systems from reproducing gender stereotypes. With that in mind, define a challenge test set for measuring the gender bias of MT systems when translating clinical reports from Basque to Spanish.
- Considering the limitations that training the MT systems at sentence level has on the MT evaluation, develop document level MT systems for translating clinical reports from Basque to Spanish.
- Being the developed systems limited to written text, extend them to speech and sign language, making them accessible to more people.
- Finally, as the systems have to be implemented in Osakidetza, and thus they are limited to translate between Basque and Spanish, collaborate with healthcare workers in the Northern Basque Country to develop MT systems for translating clinical reports between Basque and French.

Bibliography

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA. 15pp.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Barone, A. V. M., Helcl, J., Sennrich, R., Haddow, B., and Birch, A. (2017). Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*.

Bawden, R., Di Nunzio, G. M., Grozea, C., Jauregi Unanue, I., Jimeno Yepes, A., Mah, N., Martinez, D., Névéol, A., Neves, M., Oronoz, M., Perez-de Viñaspre, O., Piccardi, M., Roller, R., Siu, A., Thomas, P., Vezzani, F., Vicente Navarro, M., Wiemann, D., and Yeganova, L. (2020). Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

Biçici, E. and Yuret, D. (2015). Optimizing instance selection for statistical machine translation with feature decay algorithms. *Transactions on Audio, Speech & Language Processing*, 23(2):339–350.

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. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Etchegoyhen, T., Azpeitia, A., and Pérez, N. (2016). Exploiting a large strongly comparable corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3523–3529, Portoroz, Slovenia.

Etchegoyhen, T. and Gete, H. (2020). Handle with care: A case study in comparable corpora exploitation for neural machine translation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3799–3807, Marseille, France. European Language Resources Association.

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

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Neural Computation*, 25(2):127–144.

Graça, M., Kim, Y., Schamper, J., Khadivi, S., and Ney, H. (2019). Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy. Association for Computational Linguistics.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, UK.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

IHTSDO, I. H. T. S. D. O. (2014). *SNOMED CT Starter Guide*. Technical report, International Health Terminology Standards Development Organisation.

Joanes Etxeberri Saria V. Edizioa (2014). Donostia unibertsitate ospitaleko alta-txostenak. *Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea*.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54. Edmonton, Kanada.

Labaka, G. (2010). *EUSMT: incorporating linguistic information to SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.* PhD thesis, UPV/EHU.

Magnini, B., Altuna, B., Lavelli, A., Speranza, M., and Zanoli, R. (2020). The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In *CLiC-it*.

Mayor, A. (2007). *Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz.* PhD thesis, UPV/EHU.

McCarthy, P. M. (2005). *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity.* PhD thesis, University of Memphis, TN.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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 of the Association for Computational Linguistics*, pages 311–318. Philadelphia, AEB.

Perez-de Viñaspre, O. (2017). *Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque.* PhD thesis, UPV/EHU.

Poncelas, A., de Buy Wenniger, G. M., and Way, A. (2018a). Feature decay algorithms for neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alicante, Spain.

Poncelas, A., Popović, M., Shterionov, D., Maillette de Buy Wenniger, G., and Way, A. (2019). Combining PBSMT and NMT back-translated data for efficient NMT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 922–931, Varna, Bulgaria. INCOMA Ltd.

Poncelas, A., Shterionov, D., Way, A., de Buy Wenniger, G. M., and Passban, P. (2018b). Investigating backtranslation in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alicante, Spain.

Popović, M. (2015). chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Provilkov, I., Emelianenko, D., and Voita, E. (2020). BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

San Vicente, I. and Manterola, I. (2012). PaCo2: A fully automated tool for gathering parallel corpora from the web. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1–6, Istanbul, Turkey. European Language Resources Association (ELRA).

Sarasola, I., Salaburu, P., and Landa, J. (2015). *Hizkuntzen Arteko Corpusa (HAC)*. University of the Basque Country UPV/EHU (Euskara Institutua), Bilbao, Spain.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Barone, A. V. M., Mokry, J., and Nădejde, M. (2017). Nematus: a toolkit for neural machine translation.

Sennrich, R., Haddow, B., and Birch, A. (2015). 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, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th An-*

nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany.

Silva, C. C., Liu, C.-H., Poncelas, A., and Way, A. (2018). Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231. Cambridge, AEB.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, Montréal, Canada.

Templin, M. C. (1975). *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press, Minneapolis, MN.

Tillmann, C. and Ney, H. (2003). Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1):97–133.

Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII (Research Track)*, pages 222–232, Dublin, Ireland.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, UK.

Appendix

This appendix includes a copy of the publications related to this dissertation, in chronological order.

[**P1**] Soto X., Perez-De-Viñaspre O., Labaka G., Oronoz M. (2019) **Neural machine translation of clinical texts between long distance languages**. In *Journal of the American Medical Informatics Association* 26(12), pages 1478–1487.

[**P2**] Soto X., Perez-De-Viñaspre O., Oronoz M., Labaka G. (2019) **Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish**. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation (at MT Summit 2019)*, pages 8–18. Dublin, Ireland.

[**P3**] Soto X., Shterionov D., Poncelas A., Way A. (2020) **Selecting Back-translated Data from Multiple Sources for Improved Neural Machine Translation**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL2020)*, pages 3898–3908. (*online*).

[**P4**] Soto X., Perez-De-Viñaspre O., Labaka G., Oronoz M. (2020) **Ixamed's submission description for WMT20 Biomedical shared task: benefits and limitations of using terminologies for domain adaptation**. In *Proceedings of the 5th Conference on Machine Translation (WMT2020)*, pages 875–880. (*online*).

[**P5**] Soto X., Perez-De-Viñaspre O., Oronoz M., Labaka G. (accepted for publication) **Development of a Machine Translation system for promoting the use of a low resource language in the clinical domain: the case of Basque**. In *Natural Language Processing in Healthcare: A Special Focus on Low Resource Language. To be published.*

[**P6**] Soto X., Perez-De-Viñaspre O., Labaka G., Oronoz M. (under review) **Comparing and combining tagging with different decoding algorithms for back-translation in NMT: an analysis from a lexical diversity perspective**. Sent to *Neurocomputing*

Journal of the American Medical Informatics Association, 26(12), 2019, 1478–1487
doi: 10.1093/jamia/ocz110
Advance Access Publication Date: 23 July 2019
Research and Applications

Research and Applications

Neural machine translation of clinical texts between long distance languages

Xabier Soto,[1] Olatz Perez-de-Viñaspre,[1] Gorka Labaka,[1] and Maite Oronoz[1]

[1]Faculty of Informatics, Computer Languages and Systems, Ixa Research Group, University of the Basque Country (UPV/EHU), Donostia, Spain

Corresponding Author: Xabier Soto, MS, Faculty of Informatics, Computer Languages and Systems, Ixa Research Group, University of the Basque Country (UPV/EHU), Manuel Lardizabal 1, 20018 Donostia, Spain (xabier.soto@ehu.eus)

Received 29 January 2019; Revised 23 May 2019; Editorial Decision 26 May 2019; Accepted 31 May 2019

ABSTRACT

Objective: To analyze techniques for machine translation of electronic health records (EHRs) between long distance languages, using Basque and Spanish as a reference. We studied distinct configurations of neural machine translation systems and used different methods to overcome the lack of a bilingual corpus of clinical texts or health records in Basque and Spanish.

Materials and Methods: We trained recurrent neural networks on an out-of-domain corpus with different hyperparameter values. Subsequently, we used the optimal configuration to evaluate machine translation of EHR templates between Basque and Spanish, using manual translations of the Basque templates into Spanish as a standard. We successively added to the training corpus clinical resources, including a Spanish-Basque dictionary derived from resources built for the machine translation of the Spanish edition of SNOMED CT into Basque, artificial sentences in Spanish and Basque derived from frequently occurring relationships in SNOMED CT, and Spanish monolingual EHRs. Apart from calculating bilingual evaluation understudy (BLEU) values, we tested the performance in the clinical domain by human evaluation.

Results: We achieved slight improvements from our reference system by tuning some hyperparameters using an out-of-domain bilingual corpus, obtaining 10.67 BLEU points for Basque-to-Spanish clinical domain translation. The inclusion of clinical terminology in Spanish and Basque and the application of the back-translation technique on monolingual EHRs significantly improved the performance, obtaining 21.59 BLEU points. This was confirmed by the human evaluation performed by 2 clinicians, ranking our machine translations close to the human translations.

Discussion: We showed that, even after optimizing the hyperparameters out-of-domain, the inclusion of available resources from the clinical domain and applied methods were beneficial for the described objective, managing to obtain adequate translations of EHR templates.

Conclusion: We have developed a system which is able to properly translate health record templates from Basque to Spanish without making use of any bilingual corpus of clinical texts or health records.

Key words: neural networks, natural language processing, machine translation, long distance languages, electronic health records

INTRODUCTION AND MAIN OBJECTIVE

Our objective is to analyze different techniques for Basque-to-Spanish and Spanish-to-Basque machine translation in the clinical domain. Specifically, distinct configurations of neural machine translation (NMT) systems were tested leveraging the limited resources available for the clinical domain in Basque and Spanish.

Basque is a minoritized language, sharing a bilingual environment with the strong language Spanish. This is reflected in the

Table 1. Basque sentence translated into Spanish by a human and by different systems

Original sentence in Basque						
lipido-en	*metabolismo-aren*	*asaldura*				
lipid-GEN.PL	metabolism-GEN.SG	disorder				
"disorder of lipid metabolism"						
Manual translation into Spanish						
trastorno	*metabolismo*	*lípido-s*				
disorder	metabolism	lipid-PL				
"disorder metabolism lipids"						
Translation by Google Translate						
metabolismo	*de*	*los*	*trastorno-s*	*lipídico-s*		
metabolism	of	the.M.PL	disorder-PL	lipid-PL		
"metabolism of the lipid disorders"						
Translation by the system trained with the out-of-domain corpus						
Alteración	*de-l*	*metabolismo*	*de*	*las*	**lipides*	
disorder	of-the.M.SG	metabolism	of	the.F.PL	*lipides	
"disorder of lipides metabolism"						
Translation by the system trained including SNOMED CT terminology						
el	*trastorno*	*de-l*	*metabolismo*	*de*	*los*	*lípido-s*
the.M.SG	disorder	of-the.M.SG	metabolism	of	the.M.PL	lipid-PL
"the disorder of the lipid metabolism"						

Basque public health service, where nearly all of the health records are registered in Spanish so that any doctor can understand them. Nowadays, if any patient wants to consult their health record in Basque, it is translated on demand by human translators, the translation is given to the patient, and the public health service does not retain a copy. With a view to guaranteeing the linguistic rights of all doctors and patients, our purpose is to develop an NMT system so that Basque-speaking doctors are able to write in Basque and patients can read their health records in the language of their choice without waiting for a manual translation.

The increasing availability of electronic health records (EHRs) makes the application of advanced machine translation techniques possible. However, our main handicap is the lack of bilingual corpora for the clinical domain in Basque and Spanish. To alleviate this problem, different approaches were tried such as (1) inserting a medical bilingual dictionary to an out-of-domain corpus, (2) creating artificial sentences from the relations in SNOMED CT, (3) adding a clinical domain monolingual corpus, along with its back-translation, or (4) using the repetition of the monolingual corpus as if it was bilingual.

As a sample of the results that will be presented further, Table 1 shows an example of a sentence translated by a bilingual doctor using Google Translate,[1] our system using only out-of-domain corpora, and our system including SNOMED CT terminology.

Our main contributions are:

- The hyperparameter optimization of an NMT system dealing with long distance languages, including a morphologically rich language.
- A high-quality translation of clinical texts without an in-domain bilingual corpus of texts or records, making use of bilingual terminological resources and techniques that leverage specialized lexica and monolingual corpora.

BACKGROUND

Machine translation is defined as the process of automatically translating a text from one natural language to another. In this work, we focus on NMT, which is the result of applying the theory of Neural Networks to Machine Translation. The idea was first suggested as early as 1997,[2,3] but computational limitations did not allow it to be pursued at that time. After 15 years, the idea was recovered[4,5] with real possibilities of applying it.

Neural networks for machine translation usually rely on encoder-decoder configurations, with one neural network for each encoder and decoder. The process of training a neural network consists of making a prediction starting with some initial weights, calculating the error according to the training data, and updating the weights of the system using techniques such as backpropagation[6] until some loss function is minimized. Next, the trained model is tested with new data.

The results of NMT systems can vary depending on the architecture, number of layers, number of neurons per layer, and other configuration parameters. In order to distinguish them from the parameters (weights and bias) learned during the training process, the base configuration parameters are usually referred to as hyperparameters.

The main characteristic of NMT systems compared to previous techniques is that they act as black boxes that learn to translate without making use of any explicit linguistic or statistical information. To do this, the text in the source language is encoded into numerical values, representing word and sentence meanings as vectors, which are then decoded into sentences in the target language.

Recently, the neural approach has proven to be the most effective for machine translation when a large bilingual corpus is available,[7] making some significant improvements with the inclusion of an attention mechanism to automatically search for the most relevant words on a source sentence to be translated into the next output word[8] or using word segmentation to improve the translation of rare words.[9]

When approaching our objective, that is, building a NMT system between Basque and Spanish for the clinical domain, there are several perspectives that have to be considered, which can be divided into 3 areas: NMT between long distance languages, domain adaptation for NMT, and the handicap of performing the NMT task with no in-domain bilingual corpus. In this section, we mention the relevant works in each of these areas, although in some cases the developed techniques respond simultaneously to more than one of the described problems.

NMT between long distance languages

Spanish is a Latin-derived language that shares characteristics with other European languages, whereas Basque is a completely isolated pre-Indo-European language.

Briefly, Basque is a highly agglutinative language with a rich morphology, where words are usually created adding suffixes that mark diverse cases. The verb morphology is especially complex, including morphemes that add information about the subject, object, number, tense, aspect, etc. Furthermore, the order inside the sentences is relatively free, which makes the development of NMT systems for Basque a challenging task, particularly for evaluation purposes.

Recent work shows that better results can be obtained with NMT for Basque than with the traditional rule-based or statistical techniques.[10] Specifically, Etchegoyhen et al approach the complex morphology problem by testing different word segmentation methods, from linguistically motivated ones, to the well-known byte pair encoding (BPE) word segmentation method. For the sentence order variability, they created a second reference for the test set; they also tested different values of length-normalization and coverage penalty, based on previous work.[11] Length-normalization is a hyperparameter that compensates for the tendency of the system to choose sentences of shorter length, as the probability of a given output sentence is calculated by multiplying the probabilities of each output word. In addition, coverage penalty is used to favor sentences that cover all the words from the input sentence, with some improvements made to the attention module.[12]

Domain adaptation for NMT

Although the overall results of NMT are nowadays better than those obtained with statistical machine translation,[13] when a comparative evaluation is performed, NMT systems generate sentences with better fluency, thus sounding more natural, while statistical machine translation systems are still better in terms of accuracy.[14] Since NMT uses word embeddings to represent words, this worse accuracy is usually not a big problem provided that the generated words are similar to, or related to, the right word. In the case of the clinical domain, however, accuracy is an important aspect to preserve and some steps must be taken to improve it.

Approaches recently tested with legal domain corpora[15,16] represent a promising research area in cases in which the sentences from the training corpus are similar. They involve the same basic idea of looking for sentences similar to the input sentence before translating it, but differ in the way this sentence similarity information is used. In one case, this information is used to add the k most similar sentences to the training corpus[15]; in the other, this process is simplified using the sentence similarity scores to rescore the possible output sentences.[16]

NMT with low resources

Finally, we refer to the specific task of NMT when limited bilingual resources are available. Taking this problem to the extreme, an emerging interesting research area, known as unsupervised machine translation, is attempting to perform the NMT task without any bilingual data.

There are 2 studies[17,18] that made use of the intrinsic information contained in the word embeddings created from monolingual corpora, and then studied the best ways to relate the embedding maps created for each of the languages to be used in the translation process. Both works mark a milestone that changes the traditional paradigm that bilingual corpora are needed to perform NMT, but, as expected, they still do not obtain state-of-the-art results compared to NMT systems that make use of bilingual corpora.

There are other well established techniques that help to achieve competitive results with low resources, as in the case of transfer learning,[19] which is based on first training a system with a big enough general corpus, then fine-tuning it with a smaller corpus that can be from a specific domain; or back-translation,[20] based on including a monolingual corpus and its automatic translation to a bilingual training corpus which is similar in size. Both methods have shown to significantly improve the baseline results when some bilingual data from the domain to be tested is available (in the case of transfer learning), or a monolingual corpus of comparable size to the out-of-domain bilingual corpora is available (for back-translation).

All these techniques can be beneficial to any language pair in a domain for which there are limited bilingual resources.[21]

MATERIALS AND METHODS

System and equipment

We used the Nematus[9] NMT system, which implements the attention mechanism on RNNs and makes use of the Theano library, based on Python. Specifically, 2 different GPU servers were used: one with a Tesla K40 GPU with 12 GB of RAM, and another multi-GPU server with a Titan Xp GPU with 12 GB of RAM.

Resources

Corpora

The out-of-domain corpora used for hyperparameter optimization included a total of 4.5 million bilingual sentences. Of those, 2.3 million are a repetition of sentences from the news domain,[22] while the remaining 2.2 million sentences are from diverse domains such as administrative, web-crawling, and specialized magazines (consumerism and science). These corpora were compiled from diverse sources, such as EITB (Basque public broadcaster), Elhuyar (research foundation), and IVAP (official translation service of the Basque government). Without counting the repeated corpus, the effective data expressed in tokens were 102 000 tokens in Spanish and 72 000 tokens in Basque.

The Spanish monolingual corpus from the clinical domain was made up of real health records from the hospital of Galdakao-Usansolo consisting of 142 154 documents compiled from 2008 to 2012 with a total of 52 000 tokens. This dissociated corpus is subject to privacy agreements and is not publicly available. Table 2 summarises the data of the corpora used.

Dictionaries and other resources

Taking advantage of the resources used for the automatic translation of SNOMED CT into Basque,[23] a dictionary was built with all the created Basque terms and their corresponding Spanish entries. For many of the Spanish terms referring to a specific SNOMED CT concept, more than one possible Basque term was created. For instance, the Spanish term "lepra" (leprosy) can be translated as "legen," "legen beltz," "legenar," "negal," or "Hansen-en gaixotasun" (Hansen's disease). In total, the dictionary used for this experiment has 151 111 entries corresponding to 83 360 unique Spanish terms.

Additionally, artificial sentences were created making use of the relations specified on SNOMED CT. Specifically, the Snapshot release of the international version in RF2 format of the SNOMED

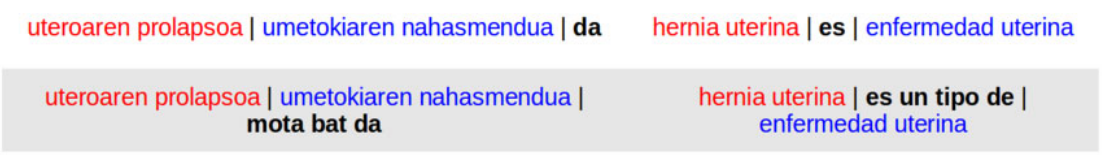

Figure 1. Sentences in Basque (left) and Spanish (right) derived from a relation in SNOMED CT.

CT delivery from July 31, 2017 was used. For the sentences to be representative, the most frequent active relations were taken into account, only considering the type of relations that appear more than 10 000 times. The most frequent active relations in the used version were "is a," "finding site," "associated morphology," and "method." As an example, a relation found in the English version is "uterine hernia" | *is a* | "disorder of uterus."

Health record templates and manual translations

For evaluating the performance of the system in the clinical domain, a total of 42 health record templates of diverse specializations written in Basque by doctors of the Donostia Hospital,[24] and their respective manual translations into Spanish carried out by a bilingual doctor, were used as reference. After aligning the sentences obtained from these EHR templates and their respective manual translations, we built a bilingual corpus consisting of 2076 sentences that were randomly ordered and divided into 1038 sentences for development (dev) and 1038 sentences for testing. Supplementary Material Table S1 shows the first 10 sentences used for evaluation in the clinical domain.

Our approach

First, we took a model NMT system between Basque and Spanish previously developed using Nematus by one of the authors (GL) and performed a hyperparameter optimization based only on the out-of-domain corpus. After this, we progressively added clinical domain resources to measure their influence on translating clinical texts. In this second part, apart from calculating BLEU scores,[25] we also carried out a human evaluation by 2 bilingual doctors who were assisted by professional translators.

NMT hyperparameter optimization

The corpus used for this part of the work was the bilingual out-of-domain one specified in the previous section, with a total of 4.5 million sentences. Specifically, 4 530 683 sentences were used for training, 1994 sentences for development, and 1678 sentences for testing. The latter ones were manually inspected for correctness prior to the testing.[10]

The starting point for this part of the work was an NMT system whose basic hyperparameters are shown in Supplementary Material Table S2.

When choosing the hyperparameters to test, various sources were consulted, but most of the hyperparameters and their possible optimal values were taken from Britz et al.[26] Supplementary Material Table S3 shows all the hyperparameters that were tried and their respective values in the same order in which they were tried.

All the experiments were carried out for both Basque-to-Spanish and Spanish-to-Basque translation directions. After comparing the results for different values of each hyperparameter, the one that achieved the highest BLEU value on the test set was chosen for the next experiment, and, only if the results were significantly different for each translation direction, a different hyperparameter value was selected for each direction.

Table 2. Summarizes the data of the corpora used

Domain	Language(s)	Documents	Sentences	Tokens
out-of-domain (news, admin., web-crawling, specialized magazines)	Basque and Spanish	–	4.5M	72M (Basque) 102M (Spanish)
clinical (EHRs)	Spanish	142 154	4.4M	52M

Table 3. Results for different hyperparameters tested out-of-domain

Translation direction	Hyperparameter update	dev BLEU	test BLEU
eu-es	Baseline	26.51	28.98
	Optimizer → Adam	26.87	28.97
	Unit type = GRU	26.87	28.97
	Beam width → 10	27.21	29.28
	Batch size → 64	27.02	**29.45**
	Embedding size → 512	26.65	28.87
es-eu	Baseline	22.95	20.26
	Optimizer → Adam	23.06	20.55
	Unit type → LSTM	23.37	20.96
	Beam width → 10	23.64	20.93
	Batch size → 64	23.05	**21.12**
	Embedding size → 512	23.09	20.42

Improvements in the test set underlined, best in bold.

Table 4. Results in the clinical domain with different training corpora

Translation direction	Training corpus	dev BLEU	test BLEU
eu-es	Baseline (Google)	6.16	5.29
	out-of-domain	10.69	10.67
	+ dictionaries	15.45	15.04
	+ artificial sentences	16.08	15.48
	+ back-translation	22.52	21.07
	+copied	23.57	**21.59**
es-eu	Baseline (Google)	2.28	2.19
	out-of-domain	9.08	8.69
	+ dictionaries	10.75	**10.44**
	+ artificial sentences	10.79	10.43

Improvements in the test set underlined, best in bold.

Evaluation in the clinical domain

First, we chose the system that achieved the best BLEU results on the out-of-domain corpus and subsequently added the following clinical domain resources to the training corpus to measure their incremental contributions to a better translation.

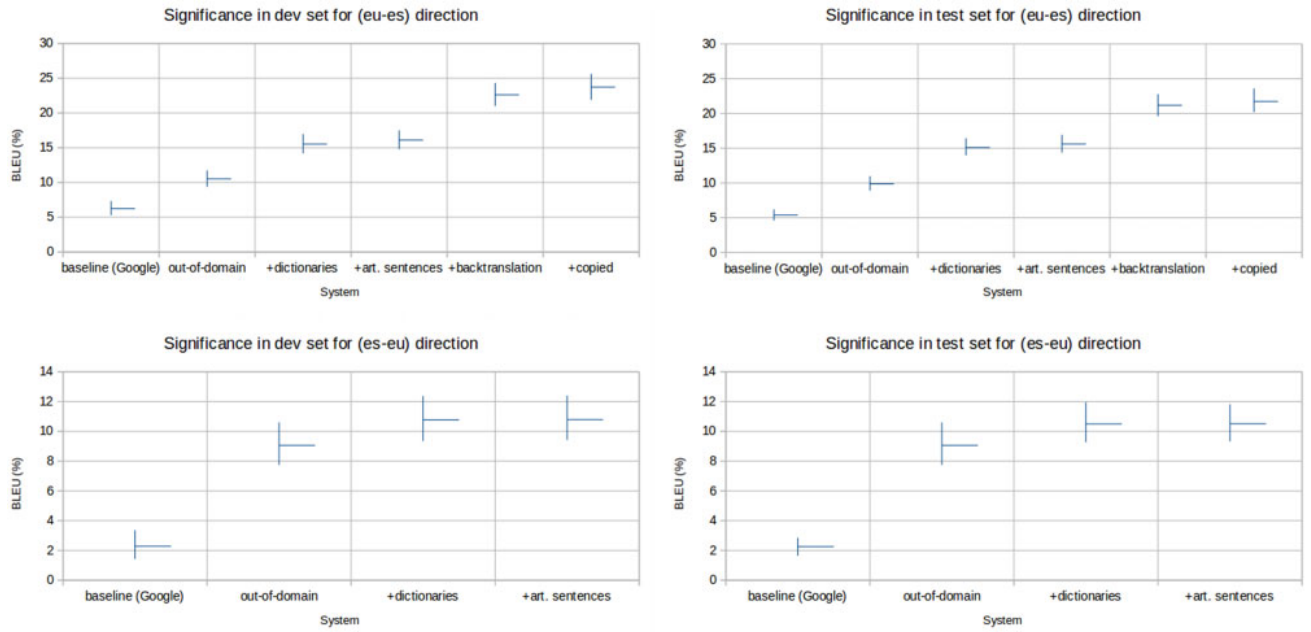

Figure 2. Results of applying bootstrap resampling on all the conducted experiments.

A dictionary from the clinical domain. For the first of these experiments, the dictionary used for the automatic translation of SNOMED CT (mentioned in the previous section) was used. For the results to be comparable with the one that only used an out-of-domain corpus, the preprocessing applied after including the dictionary was the same, consisting of tokenization, truecasing and BPE word segmentation.

Artificial sentences created from SNOMED CT. For the second of the experiments, artificial sentences created from the relations on SNOMED CT were added. The reason for adding these sentences is that NMT systems not only learn how to translate words but at the same time learn a language model from the training corpus.

To do so, we first defined 2 sentence models for each of the most frequent relations in SNOMED CT. Taking these sentence models as a reference, for each of the concepts concerning a unique pair of Basque and Spanish terms, we randomly chose one of the relations that this concept has in SNOMED CT. When doing this, we restricted the possible relations to the most frequent ones and omitted the relations with terms that had not been translated. Finally, we randomly chose one of the sentence models for this specific relation. As an example, Figure 1 shows the sentence models in Basque (left) and Spanish (right) corresponding to the previously given "is a" relation example. Equivalent terms are marked with the same color, and the meaning provided by the relation is shown in bold.

Finally, for applying the morphological inflections to the specific terms needed in some of the described sentences in Basque, a transducer was applied following the inflection rules defined in the Xuxen spelling corrector.[27] After this, a total number of 363 958 sentences were added to the corpus formed by the out-of-domain corpus and the previously added dictionary, carrying out the same preprocessing.

A monolingual corpus and its back-translation. For this part of the work, the EHRs from the Spanish monolingual corpus were used. These EHRs were first preprocessed to have one sentence in each line and then the order of the sentences of the set of EHRs was randomly changed to contribute to a better anonymization. For making the translation process faster, repeated sentences were removed from the corpus before translating it, resulting in a total of 2 023 811 sentences that were added to the previous corpus. In order to machine translate these sentences into Basque, the system specified in the first experiment was used.

A monolingual corpus as bilingual. Finally, following the work described in Currey et al,[28] we also included the same Spanish monolingual corpus and its repetition as if it were Basque, which could be beneficial for the translation of words that do not need to be translated, as in the case of drug names.

These experiments were developed for both Basque-to-Spanish and Spanish-to-Basque translation directions, except for those including the Spanish monolingual corpus, that were performed only for Basque-to-Spanish since the automatically translated corpus cannot be used as a target training corpus.[20]

RESULTS

In this section we present the results of our experiments, showing the BLEU values obtained in dev and test sets for the automatic evaluation. Basque-to-Spanish is represented in the tables as "eu-es", while Spanish-to-Basque is represented as "es-eu." As an upper bound reference for BLEU, the state-of-the-art for English-to-German machine translation is 35.0.[29] The human evaluation is performed in terms of quality and system comparison.

NMT hyperparameter optimization

Table 3 shows the results of the baseline, characterized by the hyperparameter values described in Supplementary Material Table S2,

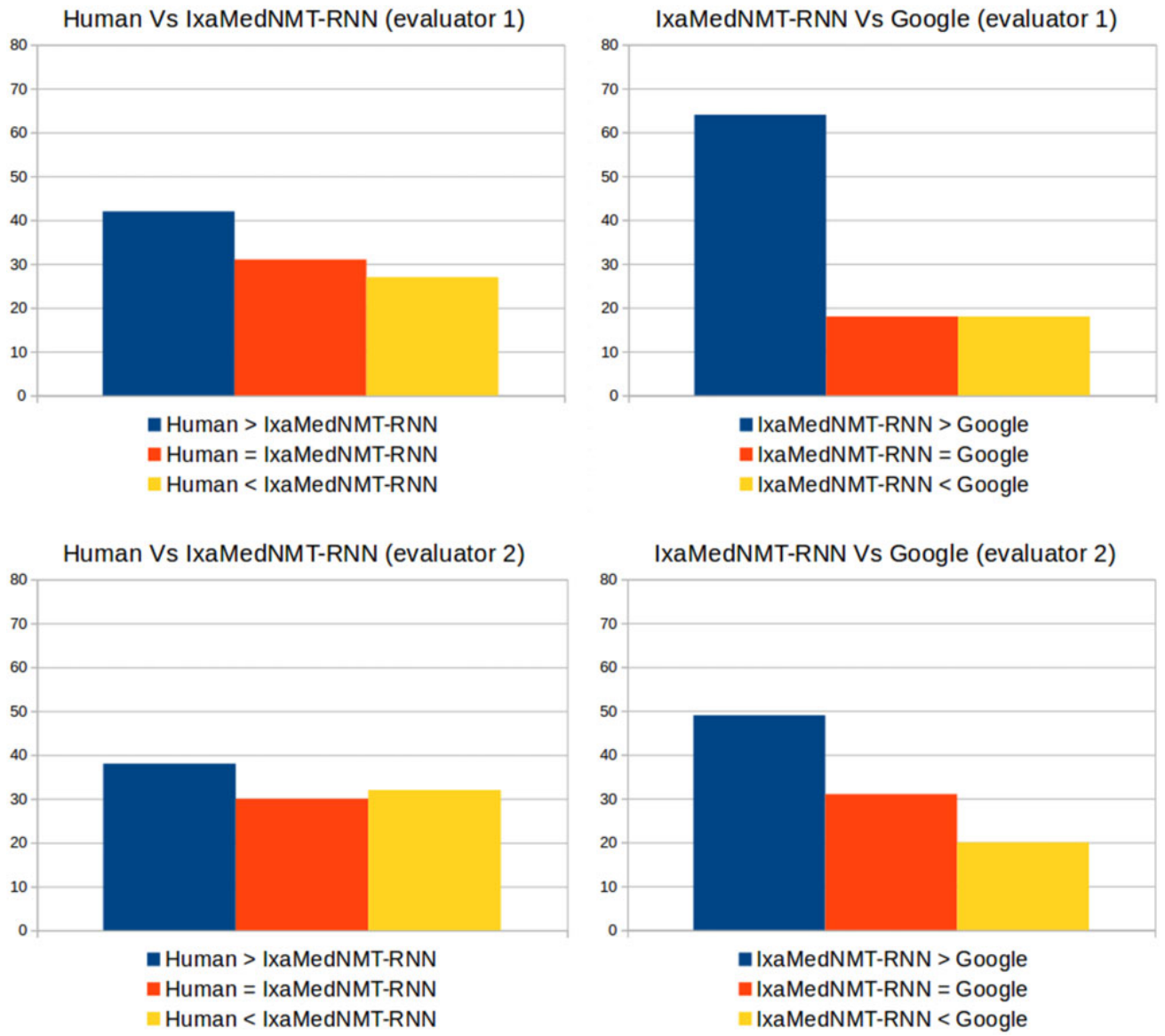

Figure 3. Comparison between human vs IxaMedNMT-RNN (left) and IxaMedNMT-RNN vs Google (right) scores given by human evaluators.

and the best results obtained with each of the hyperparameters displayed in Supplementary Material Table S3 for both translation directions in dev and test sets. Note that the results for unit type correspond to different types—Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) for each of the translation directions, as the results indicated this was the best option.

Automatic evaluation in the clinical domain

Table 4 shows the BLEU values obtained by adding resources from the clinical domain to the originally out-of-domain training corpus. As all the resources from the clinical domain were added sequentially, the "+" sign should be interpreted as an addition to the corpus corresponding to the immediate upper row.

As stated before, we only tested the inclusion of the Spanish monolingual corpus for Basque-to-Spanish translation direction. We also present the results obtained by Google Translate as a baseline, as this translator will be also taken into account in the human evaluation.

Significance

Figure 2 shows the results of applying the Moses script[30] for bootstrap resampling[31] to measure the significance of all the experiments conducted in the clinical domain. To do this, BLEU values are calculated randomly extracting 100 sentences with resampling from the corresponding set, repeating this process 1000 times, and calculating a confidence interval for the different BLEU values given a *P* value (by default, .05). As can be observed by comparing the range of BLEU values for each of the systems, only the inclusion of the dictionary and the application of the back-translation technique for Basque-to-Spanish translation direction gave improvements that could be defined as statistically significant. For both translation directions, the results of Google are significantly lower than the results of any of our systems.

Human evaluation in the clinical domain

Here we show the results of the evaluation performed by 2 bilingual doctors using the translation evaluation tool provided by TAUS[32]

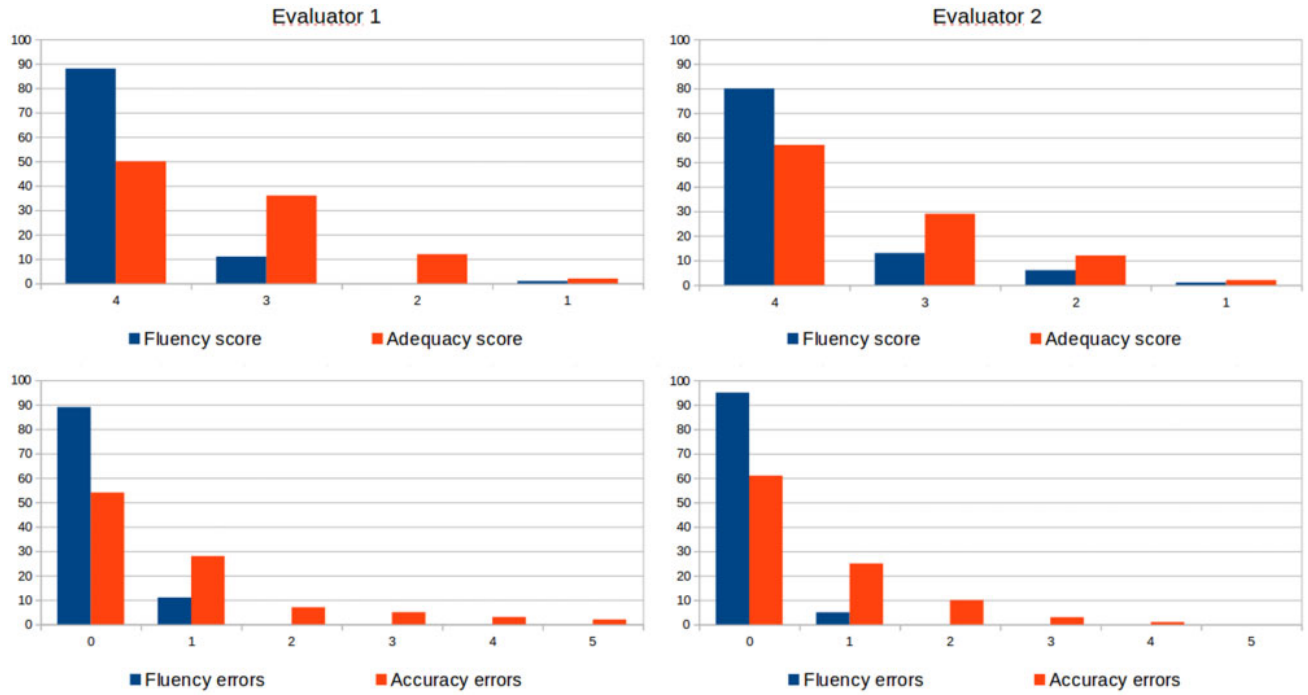

Figure 4. Fluency and adequacy scores (top); number of fluency and accuracy errors (bottom).

on 100 nonrepeated sentences randomly extracted from the test set. We carried out 2 kinds of evaluation, one for ranking the different translations made by (1) a human (reference used for the automatic evaluation); (2) the IxaMedNMT-RNN system (our best performing system in the clinical domain); and (3) Google Translate (baseline used in the previous section); and another for evaluating our IxaMedNMT-RNN system, obtaining fluency and adequacy scores (from 1 to 4), as well as the number of fluency, accuracy, terminology, style and locale convention errors in each sentence.

Figure 3 shows the comparison of the rankings given by each evaluator to the translations of the different systems. Both evaluators generally agree that the human translator is slightly better than the IxaMedNMT-RNN system, while this is much better than Google Translate. We calculated Cohen's kappa for measuring inter-annotator agreement, obtaining a 0.25 value for Human Vs IxaMedNMT-RNN comparison (fair agreement) and 0.17 for IxaMedNMT-RNN vs Google comparison (slight agreement).

In Figure 4, we provide the fluency and adequacy scores (top), together with fluency and accuracy errors (bottom) given by each evaluator. We observe that more sentences are ranked as flawless (score: 4) in terms of fluency than adequacy, while the number of accuracy errors seems to be more distributed among the translated sentences. The kappa coefficients are 0.15 (slight agreement) for fluency score and 0.65 (substantial agreement) for adequacy score. Note that for this figure and the next one, we omit the out-of-range number of errors corresponding to a sentence containing only medical analysis results, which got 20 accuracy errors according to evaluator 1 and 14 terminology errors according to evaluator 2.

Finally, Figure 5 shows the number of terminology and locale convention errors. Evaluator 1 detected 3 sentences with 1 terminology error, while evaluator 2 marked 7 sentences with 1 terminology error and 2 with 2 errors. For locale convention, evaluator one detected 2 separated errors, while evaluator 2 only marked 1 of them, the other one being a date kept in Basque format (yyyy-mm-dd). None of the evaluators detected any style error in the tested sentences.

Translation example

Figure 6 shows a clinical domain translation example.

The Google translation, although containing almost all the correct terms, loses the original meaning. Our system trained out-of-domain is unable to translate "prostata," and misses the term "tratada" (treated). The systems including the dictionary and the artificial sentences are incapable of reproducing "neoplasia," giving the inexact "tumor" and "cancer." Finally, the systems leveraging the clinical domain monolingual corpus produce flawless and adequate translations.

DISCUSSION

As Basque is a morphologically rich language, and having used the BLEU metric that counts the number of words and n-grams correctly translated, higher values are expected for Basque-to-Spanish than for Spanish-to-Basque.

When analyzing the results of hyperparameter optimization (Table 3), we observe a 0.47 point increase in the test set for Basque-to-Spanish; whereas for Spanish-to-Basque, the improvement is 0.86 points. In the case of Basque-to-Spanish, the improvement came from changing the values of beam width and batch size; whereas for Spanish-to-Basque, the results improved when changing the optimizer, unit type and batch size.

Therefore, we can conclude that the conducted experiments were mostly satisfactory (except for the embedding size) and further experiments should be carried out for both beam width and batch size.

Upon analyzing the results in the clinical domain (Table 4), we noted that all the conducted experiments improved the results,

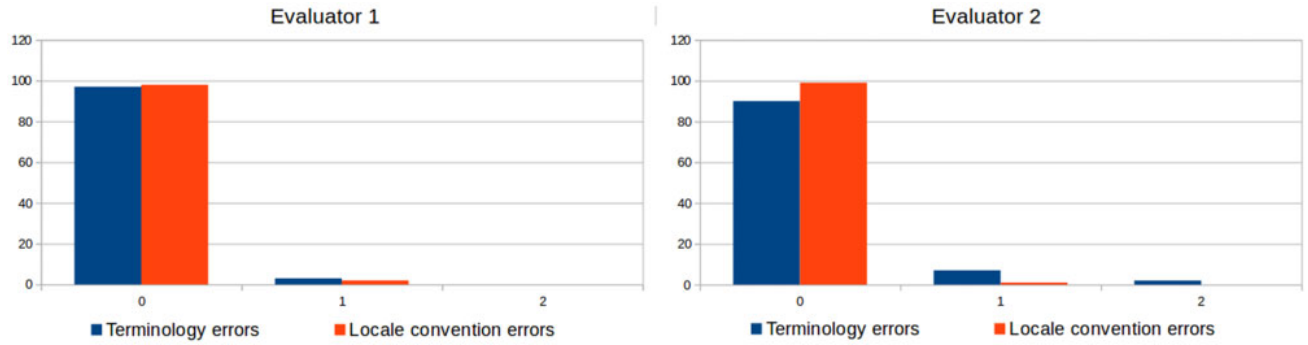

Figure 5. Number of terminology and locale convention errors.

Original sentence in Basque

2004-ko	*irail-ean*	*erradioterapia-ren*	*bidez*	*trataturiko*	*prostata-ko*	*neoplasia*
2004-GEN	september-LOC	radiotherapy-GEN	by	treated	prostate-GEN	malignancy

'malignancy of prostate treated by radiotherapy in September 2004'

Manual translation into Spanish

neoplasia	*de*	*próstata*	*tratada*	*con*	*radioterapia*	*en*	*septiembre*	*de*	*2004*
malignancy.F	of	prostate	treated.F	with	radiotherapy	in	September	of	2004

'malignancy of prostate treated by radiotherapy in September 2004'

Translation by the baseline system (Google Translator)

en	*septiembre*	*de*	*2004,*	*una*	*radioterapia*	*de*	**neoplasma*	*tratada*	*con*	*próstata*
in	September	of	2004,	a	radiotherapy.F	of	*neoplasma	treated.F	with	prostate

'In September 2004, a radiotherapy of neoplasma treated by prostate'

Translation by the system trained with the out-of-domain corpus

neoplasia	*de*	**proclasis*	*de*	*proverbio*	*en*	*septiembre*	*de*	*2004*	*a través de*	*radioterapia*
malignancy	of	*proclasis	of	proverb	in	September	of	2004	by	radiotherapy

'malignancy of proclasis of proverb in September of 2004 by radiotherapy'

Translation by the system trained including a dictionary from the clinical domain

tumor	*de*	*próstata*	*tratado*	*en*	*la*	*radioterapia*	*en*	*septiembre*	*de*	*2004*
tumor.M	of	prostate	treated.M	in	the	radiotherapy	in	September	of	2004

'tumor of prostate treated in the radiotherapy in September of 2004'

Translation by the system trained including artificial sentences created from SNOMED CT

cáncer	*de*	*próstata*	*tratada*	*en*	*septiembre*	*de*	*2004*	*a través de*	*radioterapia*
cancer.M	of	prostate	treated.F	in	September	of	2004	by	radiotherapy

'cancer of prostate treated in September of 2004 by radiotherapy'

Translation by the system trained including a monolingual corpus and its backtranslation

neoplasia	*de*	*próstata*	*tratada*	*mediante*	*radioterapia*	*en*	*septiembre*	*de*	*2004*
malignancy.F	of	prostate	treated.F	by	radiotherapy	in	September	of	2004

'malignancy of prostate treated by radiotherapy in September 2004'

Translation by the system trained including a monolingual corpus and its copy

neoplasia	*de*	*próstata*	*tratada*	*mediante*	*radioterapia*	*en*	*septiembre*	*de*	*2004*
malignancy.F	of	prostate	treated.F	by	radiotherapy	in	September	of	2004

'malignancy of prostate treated by radiotherapy in September 2004'

Figure 6. Example of a sentence translated by the different systems tested in the clinical domain.

except for the inclusion of artificial sentences that proved to be non-beneficial, especially for Spanish-to-Basque. We believe that this happened because the sentence models based on SNOMED CT relations were very simple and their syntax was already represented in the out-of-domain corpus, whereas the terminology was included in the dictionaries.

Regarding the different translation directions, it can be seen that the inclusion of each of the resources from the clinical domain has been more useful for Basque-to-Spanish. We highlight the inclusion of the dictionary, where a 4.4 BLEU points gain was achieved in the test set for Basque-to-Spanish, compared to a 1.7 points increase for Spanish-to-Basque. Given the existence of translations of SNOMED CT into many languages, a similar dictionary resource might be generated for other language pairs for which bilingual clinical corpora are lacking.

Finally, examining the results of including the different resources from the clinical domain, we conclude that the inclusion of the Spanish monolingual corpus and its translation into Basque has been the most beneficial, followed by the inclusion of the dictionary. Both results reflect that health records make use of a very specific

vocabulary and syntax, which is shown by these great improvements with the inclusion of a relatively small dictionary and a synthetic bilingual corpus formed by a monolingual corpus and its machine translation. We demonstrate that the back-translation technique, while simple, is highly effective because it helps the decoder to perform the language modeling task better.

The human evaluation confirmed these good results—ranking our system much closer to the human reference translation than to the automatic baseline system—and achieved high fluency and adequacy scores for most of the tested sentences.

For future experiments, we have to point out that even if bilingual corpora from the clinical domain becomes available, the application of the back-translation technique will also be helpful, as most of the state-of-the-art systems make use of this technique to improve their results.

CONCLUSION

We managed to optimize NMT hyperparameter values on an out-of-domain corpus, with almost 0.5 points gain in BLEU for Basque-to-Spanish, and almost 0.9 points improvement for Spanish-to-Basque from an already strong baseline.

Regarding the evaluation in the clinical domain, we point out the great improvement achieved through the technique of back-translation, with a 5.6 BLEU points gain for the tested Basque-to-Spanish translation direction. We also observe that the inclusion of the dictionary from the clinical domain has significantly improved the results, especially for Basque-to-Spanish, obtaining a 4.4 BLEU points gain. Altogether, the applied improvements have made it possible to approach the out-of-domain results, raising an acceptable result of 21.59 BLEU points for Basque-to-Spanish. These automatic evaluation results were confirmed by the human evaluation performed, showing that it is possible to develop a NMT system useful for translating clinical texts without making use of any bilingual corpus from the clinical domain.

FUNDING

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) FPI grant number BES-2017-081045, and projects DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE) and PROSA-MED (TIN2016-77820-C3-1-R, MCIU/AEI/FEDER, UE).

CONTRIBUTIONS

All authors have made substantial contributions to the conception or design of the work. XS was responsible for drafting the work; OP, GL, and MO critically revised it for important intellectual content. All authors have given final approval to the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGMENTS

We would like to thank Nora Aranberri, Idoia Arrizabalaga, Natalia Elvira, and Aitziber Etxagibel for helping us to perform the human evaluation. We would also like to thank Uxoa Iñurrieta for helping us with the glosses included in the translation example.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Google Translate. https://translate.google.com/. Accessed May 20, 2019.
2. Forcada ML, Ñeco RP. Recursive hetero-associative memories for translation In: International Work-Conference on Artificial Neural Networks Proceedings; June 4–6, 1997; Lanzarote, Canary Islands, Spain.
3. Castaño A, Casacuberta F. A connectionist approach to machine translation. In: Fifth European Conference on Speech Communication and Technology; September 22–25, 1997; Rhodes, Greece.
4. Kalchbrenner N, Blunsom P. Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing; October 18–21, 2013; Seattle, Washington, USA.
5. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems* 2014: 3104–12.
6. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 1986; 323 (6088): 533.
7. Cho K, Van Merriënboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv: 1406.1078. 2014.
8. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473. 2014.
9. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv preprint arXiv: 1508.07909. 2015.
10. Etchegoyhen T, Martínez E, Azpeitia A, *et al*. Neural machine translation of Basque. In: Proceedings of the 21st Annual Conference of the European Association for Machine Translation; Alacant, Spain; May 28–30, 2018: 139–48.
11. Wu Y, Schuster M, Chen Z, *et al*. Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv: 1609.08144. 2016.
12. Tu Z, Lu Z, Liu Y, *et al*. Modeling coverage for neural machine translation. arXiv preprint arXiv: 1601.04811. 2016.
13. Bojar O, Chatterjee R, Federmann C, *et al*. Findings of the 2016 conference on machine translation. In: ACL 2016 First Conference on Machine Translation (WMT16). The Association for Computational Linguistics; Berlin, Germany; August 11–12, 2016: 131–98.
14. Koehn P, Knowles R. Six challenges for neural machine translation. arXiv preprint arXiv: 1706.03872. 2017.
15. Gu J, Wang Y, Cho K, *et al*. Search engine guided non-parametric neural machine translation. arXiv preprint arXiv: 1705.07267. 2017.
16. Zhang J, Utiyama M, Sumita E, *et al*. Guiding neural machine translation with retrieved translation pieces. arXiv preprint arXiv: 1804.02559. 2018.
17. Artetxe M, Labaka G, Agirre E, *et al*. Unsupervised neural machine translation. arXiv preprint arXiv: 1710.11041. 2017.
18. Lample G, Denoyer L, Ranzato MA. Unsupervised machine translation using monolingual corpora only. arXiv preprint arXiv: 1711.00043. 2017.
19. Zoph B, Yuret D, May J, *et al*. Transfer learning for low-resource neural machine translation. arXiv preprint arXiv: 1604.02201. 2016.
20. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. arXiv preprint arXiv: 1511.06709. 2015.
21. Chu C, Wang R. A survey of domain adaptation for neural machine translation. arXiv preprint arXiv: 1806.00258. 2018.

22. Etchegoyhen T, Azpeitia A, Perez N. Exploiting a large strongly comparable corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation; Portorož, Slovenia; May 23–28, 2016: 3523–9.
23. Perez-de-Viñaspre O. *Automatic Medical Term Generation for a Low-Resource Language: Translation of SNOMED CT into Basque* [PhD thesis]. Donostia, Euskal Herria, University of the Basque Country; 2017.
24. Joanes Etxeberri Saria V. Edizioa Donostia Unibertsitate Ospitaleko altatxostenak. Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea, 2014.
25. Papineni K, Roukos S, Ward T, *et al*. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; Philadelphia, Pennsylvania, USA; July 6–12, 2002: 311–8.
26. Britz D, Goldie A, Luong T, *et al*. Massive exploration of neural machine translation architectures. arXiv preprint arXiv: 1703.03906, 2017.
27. Agirre E, Alegria I, Arregi X, *et al*. Xuxen: a spelling checker/corrector for basque based on two-level morphology. In: Proceedings of the Third Conference on Applied Natural Language Processing Association for Computational Linguistics; Trento, Italy; March 31 – April 3, 1992: 119–25.
28. Currey A, Barone AVM, Heafield K. Copied monolingual data improves low-resource neural machine translation. In: Proceedings of the Conference on Machine Translation (WMT). Association for Computational Linguistics; Copenhagen, Denmark; September 7–8, 2017: 148–56.
29. Edunov S, Ott M, Auli M, *et al*. Understanding back-translation at scale. arXiv preprint arXiv: 1808.09381. 2018.
30. Script Used for Measuring the Significance by Bootstrap Resampling. https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/bsbleu.py. Accessed May 20, 2019.
31. Koehn P. Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing; Barcelona, Spain; July 25–26, 2004: 388–95.
32. Webpage of the Evaluation Tool used for the Human Evaluation. https://taus.net/dqf/. Accessed May 20, 2019.

Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish

Xabier Soto, Olatz Perez-de-Viñaspre, Maite Oronoz, Gorka Labaka
Ixa Research Group, University of the Basque Country (UPV/EHU)
{xabier.soto, olatz.perezdevinaspre, maite.oronoz, gorka.labaka}@ehu.eus

Abstract

We present a method for machine translation of clinical texts without using bilingual clinical texts, leveraging the rich terminology and structure of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), which is considered the most comprehensive, multilingual clinical health care terminology collection in the world. We evaluate our method for Basque to Spanish translation, comparing the performance with and without using clinical domain resources. As a method to leverage domain-specific knowledge, we incorporate to the training corpus lexical bilingual resources previously used for the automatic translation of SNOMED CT into Basque, as well as artificial sentences created making use of the relations specified in SNOMED CT. Furthermore, we use available Electronic Health Records in Spanish for backtranslation and copying. For assessing our proposal, we use Recurrent Neural Network and Transformer architectures, and we try diverse techniques for backtranslation, using not only Neural Machine Translation but also Rule-Based and Statistical Machine Translation systems. We observe large and consistent improvements ranging from 10 to 15 BLEU points, obtaining the best automatic evaluation results using Transformer for both general architecture and backtranslation systems.

1 Introduction

The objective of this work is to study the utility of available clinical domain resources in a real use-case, which is the translation of Electronic Health Records (EHR) from Basque to Spanish. Basque is a minoritised language, also in the Basque public health service, where most of the EHRs are written in Spanish so that any doctor can understand them. With the aim of enabling Basque speaking doctors to write EHRs in Basque, we have the long-term objective of developing machine translation systems to translate clinical texts between Basque and Spanish. This work presents a method for machine translation of clinical texts from Basque to Spanish, conditioned by the current lack of clinical domain corpora in Basque.

Neural Machine Translation (NMT) has become in the past recent years the prevailing technology for machine translation, especially in the research community. Several architectures have been proposed for NMT, ranging from the initial Convolutional Neural Networks (CNN) (Kalchbrenner and Blunsom, 2013) and Recurrent Neural Networks (RNN) (Sutskever et al., 2014), to the most advanced Transformer (Vaswani et al., 2017). However, it is known that NMT systems require a large amount of training data to obtain optimal results (Koehn and Knowles, 2017), so traditional techniques as Rule-Based Machine Translation (RBMT) and Statistical Machine Translation (SMT) (Koehn et al., 2003) can be considered when the available resources are low.

One of the techniques that has become a standard to increase the available resources for NMT systems is backtranslation (Sennrich et al., 2015a), consisting in automatically translating a monolingual corpus from the target language into the

source language, and then adding both original and translated corpora to the training corpus. In our case, the availability of EHRs in Spanish enables us to improve the results for the translation of clinical texts from Basque to Spanish, also serving us as a resource for domain adaptation.

Another of our challenges is to study how to translate clinical text, which has its own characteristics differentiated from texts from other domains. Usually, the grammar of the sentences in EHRs is simplified, often omitting verbs, missing punctuation, using many acronyms and with a non-standard language more oriented to communicate between doctors than for being understood by patients. Furthermore, the main difficulty of translating clinical texts comes from the rich vocabulary used in EHRs to refer to drugs, diseases, body parts and other clinical terminology.

Regarding the language pair, our main challenge is to deal with long distance languages as Basque and Spanish, with the complexity associated with it. Specifically, we have to address the challenge of translating from a language with the characteristics of Basque. Briefly, Basque language can be described as a highly agglutinative language, with a rich morphology, where words are usually created adding diverse suffixes that mark different cases. The morphology of verbs is especially complex, including morphemes that add information about the subject, object, number, tense, aspect, etc. It is thought that the BPE word segmentation commonly used in NMT (Sennrich et al., 2015b), originally developed for avoiding the out-of-vocabulary problem, is also beneficial for the translation from morphologically rich languages as Basque.

2 Related work

Several approaches have been tried for machine translation of Basque, including Example-Based (Stroppa et al., 2006), Rule-Based (Mayor, 2007) and Statistical systems (Labaka, 2010). First works have been published for Neural Machine Translation of Basque (Etchegoyhen et al., 2018; Jauregi et al., 2018), and the first general domain commercial system for NMT between Basque and Spanish is already available online.[1]

In the NMT approach for Basque by Etchegoyhen et al. (2018), diverse morphological segmentation techniques are tested, including the aforementioned Byte Pair Encoding (BPE) (Sennrich et al., 2015b), the linguistically motivated vocabulary reduction originally proposed for Turkish (Ataman et al., 2017) and the ixaKat morphological analyser for Basque (Alegria et al., 1996; Otegi et al., 2016). They also tried character-based Machine Translation (Lee et al., 2016), obtaining the best results for translating from Basque to Spanish when applying the morphological analyser for Basque followed by BPE word segmentation to the source language corpus, and only BPE word segmentation to the target language corpus.

Regarding the clinical domain, Perez-de-Vinaspre (2017) developed a system for automatically translating the clinical terminology included in SNOMED CT (IHTSDO, 2014) into Basque. Perez-de-Vinaspre (2017) combined the use of lexical resources, transliteration of neoclassic terms, generation of nested terms and the adaptation of a RBMT system for the medical domain as backup. With respect to the translation of EHRs, the bibliography is scarce, and nowadays we can only refer to a preliminary study for translating clinical notes from English to Spanish (Liu and Cai, 2015).

Another approach for the task of translation of clinical texts is domain adaptation. Usually, when low resources for the desired domain are available, a bigger corpus from another domain is used to first train the system, which is then fine-tuned with the available in-domain corpus (Zoph et al., 2016). From another point of view, Bapna and Firat (2019) try to combine non-parametric or retrieval based approaches with NMT, looking for similarities between n-grams in the sentence to be translated and part of previously translated sentences, and then using this information for producing more accurate translations.

Concerning backtranslation, we have considered the analysis performed by Poncelas et al. (2018), where different sizes of backtranslated corpora were added to the human translated corpora used as training corpus; and regarding the techniques used for backtranslation, we follow the work by Burlot and Yvon (2019) in which they compare the performance of different SMT and NMT systems for this task.

3 Resources and methodology

As mentioned in the introduction, our main handicap is the lack of clinical domain bilingual corpora. To overcome this, we make use of available out-

[1] https://www.modela.eus/eu/itzultzailea (Accessed on April 11, 2019.)

of-domain bilingual corpora, automatically created clinical terminology in Basque (Perez-de-Viñaspre, 2017), artificial sentences formed based on the relations specified in SNOMED CT, and EHRs in Spanish that are used for backtranslation (Sennrich et al., 2015a) and copying (Currey et al., 2017).

For evaluation in the clinical domain, we use EHR templates in Basque published with academic purposes (Joanes Etxeberri Saria V. Edizioa, 2014), together with their manual translations into Spanish performed by a bilingual doctor.

In the following, we present the details of each of the resources and explain how they were used in this work.

3.1 Out-of-domain corpora

As a basis for our work, we use a large bilingual corpus formed by 4.5 million sentences, where 2.3 million sentences are a repetition of a corpus from the news domain (Etchegoyhen et al., 2016), and the remaining 2.2 million sentences are from diverse domains such as administrative, web-crawling and specialised magazines (consumerism and science). These corpora were compiled from sources such as EITB (Basque public broadcaster), Elhuyar (research foundation) and IVAP (Basque institute of public administration).

3.2 Clinical terminology

As a first step for improving the translation of clinical texts, we built a dictionary with all the Basque terms and their corresponding Spanish entries used for the automatic translation of SNOMED CT into Basque (Perez-de-Viñaspre, 2017). These terms were compiled from different sources such as Euskalterm, Elhuyar Science and Technology dictionary, UPV/EHU human anatomy atlas and nursery dictionary, International Classification of Diseases dictionary and a health administration related dictionary. As this work corresponds to a first approach of developing a Basque version of SNOMED CT, more than a possible Basque term was created for each entry in Spanish. Altogether, we use 151,111 Basque terms corresponding to 83,360 unique Spanish terms. We think that the fact of having more than one possible Basque term for each Spanish entry helps us to improve the coverage of the system for translating from Basque to Spanish. As a sample, Table 1 shows the first 10 clinical terms included as training corpus.

3.3 Artificial sentences

While including clinical terms in our system helps us to approach the rich terminology characteristic of clinical notes, we think that including these same terms in the form of sentences could be more suitable to the task of translating sentences from EHRs. For doing this, we leverage the structured form of SNOMED CT, using the relations specified in it to create simple artificial sentences that could be more similar to the real sentences included in EHRs.

Specifically, the Snapshot release of the international version on RF2 format of the SNOMED CT delivery from 31st July 2017 was used. For the sentences to be representative, the most frequent active relations were taken into account, only considering the type of relations that appeared more than 10,000 times. The most frequent active relations in the used version were "is a", "finding site", "associated morphology" and "method".

For creating the artificial sentences, we first defined two sentence models for each of the most

Basque term	Spanish term	English gloss
organo kopulatzaile	órgano copulador	*copulatory organ*
dionisiako	dionisiaco	*Dionysian*
desfile	desfile	*parade*
begi-miiasia	miasis ocular	*ophthalmic myiasis*
ahoko kandidiasi	candidiasis oral	*oral candidiasis*
wolfram	wolframio	*Tungsten*
W	wolframio	*Tungsten*
zergari	recaudador	*collector*
jasotzaile	recaudador	*collector*
biltzaile	recaudador	*collector*

Table 1: First 10 clinical terms included as training corpus.

frequent relations in SNOMED CT. Taking these sentence models as a reference, for each of the concepts concerning a unique pair of Basque and Spanish terms, we randomly chose one of the relations that this concept has in SNOMED CT. When doing this, we restricted the possible relations to the most frequent ones and omitted the relations with terms that were not available in both languages. Finally, we randomly chose one of the two sentence models for this specific relation.

Considering the agglutinative character of Basque language, some of the created sentences needed the application of morphological inflections to the specific terms included in the artificial sentences. For this task, a transducer was applied following the inflection rules defined in the Xuxen spelling corrector (Agirre et al., 1992). In total, 363,958 sentences were created. As a sample, Table 2 shows the first 10 artificial sentences created with this method, separating different terms and relations with 'I', giving the same superscript number to equivalent terms, and marking the terms that define the relations in bold.

3.4 EHRs in Spanish

Finally, as a main contribution to the translation of clinical texts, we make use of available EHRs in Spanish. This corpus is made up of real health records from the hospital of Galdakao-Usansolo consisting of 142,154 documents compiled from

Basque sentence	**Spanish sentence**
umetokiaren prolapsoa[1] I emakumezkoaren prolapso genitala, zehaztugabea[2] I **da** *uterine prolapse*[1] I ***is a*** I *prolapse of female genital organs, undefined*[2]	prolapso uterino[1] I **es** I prolapso de los órganos genitales femeninos[2] *uterine prolapse*[1] I ***is a*** I *prolapse of female genital organs*[2]
umetokiaren prolapsoa[1] I uteroa[2] I **-n gertatzen da** *uterine prolapse*[1] I ***occurs in*** I *uterus*[2]	descenso uterino[1] I **ocurre en** I estructura uterina[2] *descensus uteri*[1] I ***occurs in*** I *uterine structure*[2]
umetokiaren prolapsoa[1] I uteroaren egitura[2] I **-n aurkitzen da** *uterine prolapse*[1] I ***is found in*** I *uterine structure*[2]	hernia uterina[1] I **se encuentra en** I estructura uterina[2] *uterine hernia*[1] I ***is found in*** I *uterine structure*[2]
uteroaren prolapsoa[1] I emakumezkoaren prolapso genitala, zehaztugabea[2] I **da** *uterine prolapse*[1] I ***is a*** I *prolapse of female genital organs, undefined*[2]	prolapso uterino[1] I **es** I prolapso genital[2] *uterine prolapse*[1] I ***is a*** I *genital prolapse*[2]
uteroaren prolapsoa[1] I umetokiko trastorno ez-inflamatorioa, zehaztugabea[2] I **mota bat da** *uterine prolapse*[1] I ***is a type of*** I *noninflammatory uterine disorder, undefined*[2]	descenso uterino[1] I **es un tipo de** I trastorno uterino[2] *descensus uteri*[1] I ***is a type of*** I *uterine disorder*[2]
uteroaren prolapsoa[1] I umetokiaren nahasmendua[2] I **da** *uterine prolapse*[1] I ***is a*** I *uterine disorder*[2]	hernia uterina[1] I **es** I enfermedad uterina[2] *uterine hernia*[1] I ***is a*** I *uterine disease*[2]
zakilaren inflamazioa[1] I zakil[2] I **-ean gertatzen da** *inflammation of penis*[1] I ***occurs in*** I *penis*[2]	inflamación del pene[1] I **ocurre en** I estructura de pene[2] *inflammation of penis*[1] I ***occurs in*** I *penis structure*[2]
zakilaren inflamazioa[1] I zakilaren egitura[2] I **-n aurkitzen da** *inflammation of penis*[1] I ***is found in*** I *penis structure*[2]	trastorno inflamatorio del pene[1] I **se encuentra en** I pene[2] *inflammatory disorder of penis*[1] I ***is found in*** I *penis*[2]
zakilaren hantura[1] I zakilaren gaitza[2] I **da** *inflammation of penis*[1] I ***is a*** I *disorder of penis*[2]	inflamación del pene[1] I **es** I enfermedad peniana[2] *inflammation of penis*[1] I ***is a*** I *disorder of penis*[2]
zakilaren hantura[1] I zakilaren gaitz[2] I **mota bat da** *inflammation of penis*[1] I ***is a type of*** I *disorder of penis*[2]	trastorno inflamatorio del pene[1] I **es un tipo de** I enfermedad peniana[2] *inflammatory disorder of penis*[1] I ***is a type of*** I *disorder of penis*[2]

Table 2: First 10 artificial sentences created from relations in SNOMED CT.

2008 to 2012. Due to privacy agreements, this dissociated corpus is not publicly available.

These documents were first preprocessed to have one sentence in each line, and then the order of the sentences was randomly changed to contribute to a better anonymisation. For making the translation process faster, repeated sentences were removed from the corpus before translating it, resulting in a total of 2,023,811 sentences.

This corpus was added twice to the training corpus, once by applying different backtranslation techniques, and the other by simply using the same corpus in Spanish as if it were Basque (Currey et al., 2017), which we think could be beneficial for the translation of words that do not need to be translated, as it is the case of drug names. This way, from the total number of sentences used for training the corpus based systems developed for translation of clinical texts (9,093,374), around half of them correspond to out-of-domain sentences (4,530,683), and the other half come from diverse clinical domain sources (4,562,691).

Table 3 summarises the numbers of the training corpora. All corpora was tokenised and truecased using the utilities of Nematus (Sennrich et al., 2017) if they were to be used for corpus based systems. For NMT experiments, BPE word segmentation was performed using subword-nmt[2], applying 90,000 merge operations on the joint bilingual corpora. The number of tokens in Basque for the backtranslated EHRs correspond to the backtranslation performed with the shallow RNN.

3.5 EHR templates in Basque and their manual translations into Spanish

For evaluating the task of translating clinical texts, we used 42 EHR templates of diverse specializations written in Basque by doctors of the Donostia Hospital, and their respective manual translations into Spanish carried out by a bilingual doctor. We manually aligned the sentences from these templates with their respective translations, building a bilingual corpus of 2,076 sentences. These sentences were randomly ordered and further divided into 1,038 sentences for development purposes and 1,038 sentences for test purposes.

We highlight that the sentences used for evaluation in the clinical domain come from diverse specializations, which we expect to be mirrored in a more diverse set of development and test corpora. Furthermore, from the 1,038 sentences from the test set, 826 are non-repeated, corresponding the most repeated ones to short sentences relating to EHR section titles. As a sample, Table 4 shows the first 10 sentences used for evaluation in the clinical domain.

4 Experiments

We test our method through two types of experiments, one regarding different NMT architectures, and the other referring to different systems used for backtranslation. All the experiments concerning NMT systems were performed on Titan XP GPUs, using only one for training the shallow RNN, and two for the deep RNN and the Transformer.

4.1 Architectures

First, we test the performance of several neural architectures, trying a shallow RNN as an easily reproducible system, a Transformer (Vaswani et al., 2017) architecture as state-of-the-art performing system, and a deep RNN (Barone et al., 2017) as a fairer comparison to Transformer.

We develop two systems for each architecture, one trained only with out-of-domain corpora, and another trained with all the available resources, including the ones from the clinical domain. For this part of the work, the backtranslation of the available EHRs in Spanish was performed by the shallow RNN.

We evaluate the performance of all the systems in the clinical domain, using the EHR templates in

[2] https://github.com/rsennrich/subword-nmt (Accessed on April 11, 2019.)

Domain	Type	Sentences	Tokens
out-of-domain	Diverse sentences	4.5 million	73 million (Basque) / 102 million (Spanish)
clinical domain	Terms	151,111	271,248 (Basque) / 257,641 (Spanish)
	Artificial sentences	363,958	3.1 million (Basque) / 4.1 million (Spanish)
	Backtranslated EHRs	2 million	26 million (Basque) / 33 million (Spanish)
	Copied EHRs	2 million	33 million

Table 3: Numbers of the training corpora.

Basque sentence	**Spanish sentence**
tratamendua *therapy*	tratamiento *therapy*
abortuak: 1 *aborta 1*	abortos 1 *aborta 1*
lehenengo sintomatologia *first symptomatology*	primera sintomatología *first symptomatology*
fibrinolisiaren ondoren egoera klinikoa ez da askorik aldatu *clinical status does not change much after fibrinolysis*	la situación clínica después de la fibrinólisis no cambia sustancialmente *clinical status after fibrinolysis does not change substantially*
hipertentsioaren aurkako tratamenduarekin hasi da, tentsioak neurri egokian mantenduz; hipergluzemiarako joera antzeman da egonaldian *he/she started the treatment for hypertension, keeping tensions at the right level; a tendency to hyperglycemia is observed during the stay*	al mismo tiempo tratamientopara normalizar la HTA, hiperglucemia y dislipemia *at the same time treatmentfor* normalising HBP, hyperglycemia and dislipemia**
ebakuntza aurreko azterketa normala izan ostean, 2012-08-20an operazioa egin da *after the preoperative examination being normal, the operation is done on 2012-08-20*	tras ser normal la exploración preoperatoria se opera el 20-08-2012, practicándose: *after the preoperative exploration being normal he/she is operated on 2012-08-20, by practising:*
Dismetriarik ez *No dysmetria*	no dismetría *no dysmetria*
miaketa oftalmologikoa normala *normal ophthalmic exploration*	examen oftalmológico normal *normal ophthalmic examination*
EKG: erritmo sinusala, 103 tau/min *ECG: sinus rhythm, 103 beat/min*	EKG-ritmo sinusal 103/minuto *ECG-sinus rhythm, 103/min*
ez du botaka egin *he/she has not vomited*	no vómitos *no vomits*

Table 4: First 10 sentences used for evaluation in the clinical domain.

Basque and their manual translations into Spanish specified in the previous section.

A description of the tested architectures is given in the following lines.

Shallow RNN: As a simple RNN, we use a model developed with the old version of Nematus (Sennrich et al., 2017), making use of the Theano framework. Specifically, we use 1 layer (bidirectional for the encoder) of 1024 GRU (Cho et al., 2014) units, with a embedding-size of 500, a batch-size of 64 and using Adam (Kingma and Ba, 2014) as optimisation method. For decoding, we use a beam-width of 10 for all the experiments. Some of the values of these hyperparameters were optimised with the out-of-domain corpus, and subsequently used in the other architectures.

Deep RNN: As a more advanced RNN, we select the system developed by Barone et al. (2017), included in a more recent work in which linguistic abilities of diverse NMT systems were tested (Tang et al., 2018).

From the different variants presented in Barone et al. (2017), we use the one that obtained the best reported results, whose configuration parameters are public.[3]

Transformer: As a state-of-the-art NMT system, we choose the Transformer implementation in Pytorch by OpenNMT (Klein et al., 2017). We use the recommended hyperparameters,[4] except for the number of GPUs and batch-size, that were

[3] `https://github.com/Avmb/deep-nmt-architectures/blob/master/configs/bideep-bideep-rGRU-large/config.sh` (Accessed on April 11, 2019.)

[4] `http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model-do-you-support-multi-gpu` (Accessed on April 11, 2019.)

halved to meet our hardware capabilities.

4.2 Backtranslation systems

After trying different architectures, we select the one that obtains the best automatic evaluation results in the clinical domain and change the way the backtranslation is performed. For that, we compare the shallow RNN architecture with the one that gets the best results in the clinical domain, and also try RBMT and SMT systems to translate the EHRs in Spanish into Basque.

For training the corpus based systems in the Spanish-to-Basque translation direction, we use the out-of-domain corpus and the dictionaries including clinical terminology. The resulting synthetic corpus is added together with the artificial sentences and the copied monolingual corpus, and the performance of the systems is tested in the clinical domain.

Shallow RNN: For this experiment we use the same shallow RNN architecture specified in the previous section, just changing the translation direction. Note that, due to an error in the preprocessing, the BPE word segmentation was performed for 45,000 steps in each language corpus, instead of 90,000 times in the joint corpora. We do not expect for this error to have significant influence on the final results.

Transformer: We train the Transformer system in the Spanish-to-Basque translation direction with the same hyperparameters specified in the previous section. Following the work by Edunov et al. (2018), we perform the translation by unrestricted random sampling, which is proved to obtain better results than restricted random sampling or traditional beam search when applied to backtranslation.

RBMT: For this part of the work, we try Matxin (Mayor, 2007), a Rule-Based system for Spanish-to-Basque Machine Translation, adapted to the biomedical domain by the inclusion of dictionaries. In this case, we translate the EHRs in Spanish before truecasing, so when removing the repeated sentences from the corpora the number of sentences is not exactly the same as for the monolingual corpus translated with corpus based systems (2,036,165 instead of 2,023,811).

SMT: Finally, we try Moses (Koehn et al., 2007) as a statistical system, adapted to the biomedical domain. We use default parametrisation with MGIZA for word alignment, a "msd-bidirectional-fe" lexicalised reordering model and a KenLM (Heafield, 2011) 5-gram target language model. The weights for the different components were adjusted to optimise BLEU using Minimum Error Rate Training (MERT) with an n-best list of size 100.

5 Results and discussion

In this section we show and discuss the automatic evaluation results of the experiments carried out with different architectures and backtranslation systems. In both cases, we calculate BLEU (Papineni et al., 2002) in development and test sets using the multi-bleu script included in Moses.[5]

5.1 Architectures

Table 5 shows the results of the tested architectures in two variants: 1) trained only with out-of-domain corpora, and 2) including all the clinical domain resources. We observe large and consistent improvements when adding in-domain data to each of the tested architectures. Surprisingly, the deep RNN obtains lower results than the shallow RNN, especially comparing the systems trained out-of-domain, which can be an overfitting issue. We also think that the previous optimisation with the out-of-domain corpus of some of the hyperparameters of the shallow RNN can be a reason for its good results, comparable with Transformer regarding the systems trained only with out-of-domain corpora, and similar to deep RNN when adding the clinical domain resources.

	dev	test
Shallow RNN (out-of-domain)	10.69	10.67
Shallow RNN (+in-domain)	23.57	21.59
Deep RNN (out-of-domain)	7.23	5.91
Deep RNN (+in-domain)	23.01	20.74
Transformer (out-of-domain)	10.92	10.55
Transformer (+in-domain)	**26.67**	**24.44**

Table 5: BLEU values (Basque-to-Spanish) for different architectures using a shallow RNN for backtranslation.

However, if we compare the results of the different architectures trained with all the available re-

[5] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl` (Accessed on April 11, 2019.)

sources, we see that Transformer outperforms both RNNs by around 3 BLEU points in each evaluation set. Thus, we can say that the Transformer architecture is the optimal for our task of translating clinical texts from Basque to Spanish.

5.2 Backtranslation systems

After determining which is the best general architecture for our task, we compare the results of different backtranslation systems. First, we evaluate the performance of the systems used to translate the available EHRs in Spanish into Basque, using as a reference the same datasets employed for evaluating the different architectures. Table 6 shows the results of the tested backtranslation systems.

	dev	test
RBMT$_{bt}$	8.56	7.03
SMT$_{bt}$	10.30	8.75
Shallow RNN$_{bt}$	10.75	10.44
Transformer$_{bt}$	**11.30**	**12.04**

Table 6: BLEU values for different backtranslation systems (Spanish-to-Basque).

We observe that the values obtained with NMT systems are similar to the ones obtained in the other direction with the system trained out-of-domain, which is logical since we only added the dictionaries for training the backtranslation systems. The results of SMT are also similar, with a slightly lower score in the test set. The results for RBMT are even lower, which can be because BLEU underestimates the results of RBMT systems in general.

Finally, we present in Table 7 the results in the clinical domain of the systems trained with the best performing architecture (Transformer) using all the training corpora, changing the method used for backtranslating the EHRs in Spanish.

	dev	test
RBMT	22.98	21.91
SMT	22.78	21.43
Shallow RNN	26.67	24.44
Transformer	**27.70**	**25.61**

Table 7: BLEU values (Basque-to-Spanish) for Transformer architecture using different backtranslation systems.

We notice that using Transformer for backtranslation obtains the best results, gaining more than 1 BLEU point comparing with the same Transformer architecture using a shallow RNN for backtranslation. The results for RBMT and SMT are lower, but comparing to the BLEU values for the backtranslation systems (Table 6), we observe that in this case the results using RBMT are slightly better than the ones with SMT. Apart from the aforementioned possible underestimation of RBMT systems when calculating BLEU, we think that this could be because the RBMT system can translate words that corpus based systems cannot translate, adding more variability to the source language corpus.

5.3 Ensemble of best models

After evaluating the performance of different architectures and backtranslation systems, we evaluate the performance of an ensemble of the 3 systems obtaining highest BLEU values in the development set, which in this case correspond to 3 different models of the Transformer architecture, using Transformer as backtranslation system, saved after different number of iterations. Specifically, the models evaluated for the ensemble are those saved after 90,000, 160,000 and 180,000 iterations, obtaining 27.56 BLEU points with the first two models, and 27.70 BLEU points with the last one. Table 8 shows the results of the ensemble system, which we name IxaMedNMT-Transformer. We observe gains of 0.33 BLEU points in the development set and 0.11 BLEU points in the test set, comparing to the results of the single model that obtained the highest BLEU value in the development set.

	dev	test
IxaMedNMT-Transformer	28.03	25.72

Table 8: BLEU values (Basque-to-Spanish) for an ensemble of the best performing systems.

5.4 Translation example and error analysis

Finally, Figure 1 shows an example of a translation performed by the ensemble system whose BLEU values were shown in Table 8, along with the original sentence in Basque and the manual translation into Spanish used as a reference.

We observe that the generated translation is almost equivalent to the human translation, with only slight differences in some of the words (presents/with, complete/wide, stenoses/obstructs, part/region, etc.), but without changing the overall meaning of the original sentence in Basque.

Original sentence in Basque

azaleko izter-arteria-k buxadura zabala du, baina iragazkor dago Hunter-en eremu-raino,
superficial femoral artery-ERG obstruction wide has, but permeable is Hunter-GEN region-ALL,
'the superficial femoral artery has a wide obstruction, but it is permeable up to the Hunter region,...

bertan buxatu eta 3. eremu popliteo-an berriz ere iragazkor bihurtzen da.
there obstruct and 3rd region popliteal-LOC again also permeable becoming is.
it is obstructed there and becomes permeable again in the 3rd popliteal region.'

Manual translation into Spanish

arteria femoral superficial con estenosis amplia pero permeable hasta la zona de Hunter
artery femoral superficial with stenosis wide but permeable up-to the region of Hunter
'superficial femoral artery with wide obstruction but permeable up to the Hunter region...

donde se estenosa, y en la zona 3 poplítea se vuelve otra vez permeable.
where it stenoses, and in the region 3 popliteal it becomes another.F time permeable.

where it stenoses and becomes permeable again in the popliteal region 3.'

Translation by the IxaMedNMT-Transformer system

la arteria femoral superficial presenta una oclusión completa que se encuentra permeable hasta el
the artery femoral superficial presents a occlusion complete which is found permeable up-to the
'the superficial femoral artery presents a complete occlusion which is permeable up to the...

área de Hunter, donde se obstruye y se vuelve permeable en la 3ª porción poplítea.
region of Hunter, where it obstructs and it becomes permeable in the 3rd portion popliteal.

Hunter region, where it is obstructed and becomes permeable in the 3rd popliteal portion.'

Figure 1: Translation example by the IxaMedNMT-Transformer system, along with the original sentence in Basque and the manual translation into Spanish.

In a fast overview of the whole of the sentences translated from the development set, we have observed that for some of the long sentences, the translation ended abruptly without translating a few of the last words. We have tried to scale down the beam-width from 10 (optimised for the shallow RNN, kept in other architectures for fair comparison) to the default value of 5 to reduce the probability of generating the end-of-sentence token sooner than necessary, but the BLEU values in the development set did not improve as expected. We plan to test diverse values of length-normalisation and coverage-penalty coefficients to try to overcome this problem.

This phenomenon occurred especially in sentences with a lot of punctuation marks, usually containing a list of symptoms, diseases or drugs. Regarding the translation of rare words, like in this case drug names, we have observed very few errors where part of the word was not translated correctly due to the BPE word segmentation. In the future, we intend to perform a thorough analysis of the different types of errors encountered in the generated translations, with the aim of developing possible solutions to them.

6 Conclusions and future work

We have showed that it is possible to translate clinical texts from Basque to Spanish without clinical domain bilingual corpora. We have leveraged previous work in translation of clinical terminology into Basque (Perez-de-Viñaspre, 2017), described a method for creating artificial sentences based on SNOMED CT relations, and made use of available EHRs in Spanish. Given the multilinguality and rich structure of SNOMED CT, similar dictionaries and artificial sentences might be generated for other language pairs for which bilingual clinical corpora are not available.

Furthermore, we have tested our method with different NMT architectures and using diverse systems for backtranslation, including rule-based and statistical systems. We obtained the best results using Transformer for both general architecture and backtranslation systems, achieving 28 BLEU points in the development set through checkpoint ensembling, and showing a translation example.

We leave as future work the human evaluation of the best performing system, with the possibility of improving the corpora used for training and evaluation.

Acknowledgements: This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) FPI grant number BES-2017-081045, and projects BigKnowledge (BBVA foundation grant 2018), DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE) and PROSA-MED (TIN2016-77820-C3-1-R, MCIU/AEI/FEDER, UE). We thank Uxoa Iñurrieta for helping us with the glosses.

References

Agirre, Eneko, Inaki Alegria, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola, and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on Two-Level morphology. In *Proceedings of the third conference on Applied natural language processing*, 119–125.

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.

Bapna, Ankur, and Orhan Firat. 2019. Non-Parametric Adaptation for Neural Machine Translation. *arXiv preprint arXiv:1903.00058*

Barone, Antonio Valerio Miceli, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. *arXiv preprint arXiv:1707.07631*

Burlot, Franck, and François Yvon. 2019. Using Monolingual Data in Neural Machine Translation: a Systematic Study. *arXiv preprint arXiv:1903.11437*

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, 148–156.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*

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)*, Porotoroz, Slovenia.

Etchegoyhen, Thierry, Eva Martínez, Andoni Azpeitia, Gorka Labaka, Iñaki Alegria, Itziar Cortes, Amaia Jauregi, Igor Ellakuria, Maite Martin and Eusebi Calonge. 2018. Neural Machine Translation of Basque. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 139–148.

Heafield, Kenneth. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187–197.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

International Health Terminology Standards Development Organisation IHTSDO. 2014. *SNOMED CT Starter Guide*. Technical report, International Health Terminology Standards Development Organisation

Jauregi, Inigo, Lierni Garmendia, Ehsan Zare, and Massimo Piccardi. 2018. English-Basque statistical and neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 880–885.

Joanes Etxeberri Saria V. Edizioa. 2014. *Donostia Unibertsitate Ospitaleko alta-txostenak*. Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea

Kalchbrenner, Nal, and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1700–1709.

Kingma, Diederik P., and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*

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, 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 on Human Language Technology-Volume 1*, 48–54.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, 177–180.

Koehn, Philipp, and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*

Labaka, Gorka. 2010. *EUSMT: incorporating linguistic information into SMT for a morphologically rich language. Its use in SMT-RBMT-EBMT hybridation.*. PhD thesis, University of the Basque Country, Donostia, Euskal Herria.

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

Liu, Weisong, and Shu Cai. 2015. Translating electronic health record notes from English to Spanish: A preliminary study. *Proceedings of BioNLP 15*, 139–148.

Mayor, Aingeru. 2007. *Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz.*. PhD thesis, University of the Basque Country, Donostia, Euskal Herria.

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

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*, 311–318.

Perez-de-Viñaspre, Olatz. 2017. *Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque*. PhD thesis, University of the Basque Country, Donostia, Euskal Herria.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation. *arXiv preprint arXiv:1804.06189*

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nădejde. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*

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*, Cambridge, MA USA, 232–241.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Tang, Gongbo, Mathias Müller, Annette Rios, and Rico Sennrich. 2018. Why self-attention? a targeted evaluation of neural machine translation architectures. *arXiv preprint arXiv:1808.08946*

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*, 5998–6008.

Zeiler, Matthew D. 2012. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*

Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation

Xabier Soto,[1] Dimitar Shterionov,[2] Alberto Poncelas,[2] and Andy Way[2]
[1]Ixa NLP Group, HiTZ Center, University of the Basque Country (UPV/EHU)
[2]ADAPT Centre, School of Computing, Dublin City University
[1]xabier.soto@ehu.eus
[2]{firstname.lastname}@adaptcentre.ie

Abstract

Machine translation (MT) has benefited from using synthetic training data originating from translating monolingual corpora, a technique known as backtranslation. Combining backtranslated data from different sources has led to better results than when using such data in isolation. In this work we analyse the impact that data translated with rule-based, phrase-based statistical and neural MT systems has on new MT systems. We use a real-world low-resource use-case (Basque-to-Spanish in the clinical domain) as well as a high-resource language pair (German-to-English) to test different scenarios with backtranslation and employ data selection to optimise the synthetic corpora. We exploit different data selection strategies in order to reduce the amount of data used, while at the same time maintaining high-quality MT systems. We further tune the data selection method by taking into account the quality of the MT systems used for backtranslation and lexical diversity of the resulting corpora. Our experiments show that incorporating backtranslated data from different sources can be beneficial, and that availing of data selection can yield improved performance.

1 Introduction

The use of supplementary backtranslated text has led to improved results in several tasks such as automatic post-editing (Junczys-Dowmunt and Grundkiewicz, 2016; Hokamp, 2017), machine translation (MT) (Sennrich et al., 2016a; Poncelas et al., 2018b), and quality estimation (Yankovskaya et al., 2019). Backtranslated text is a translation of a monolingual corpus in the target language (L2) into the source language (L1) via an already existing MT system, so that the aligned monolingual corpus and its translation can form an L1–L2 parallel corpus. This corpus of synthetic parallel data can then be used for training, typically alongside authentic human-translated data. For MT, backtranslation has become a standard approach to improving the performance of systems when additional monolingual data in the target language is available.

While Sennrich et al. (2016a) show that any form of source-side data (even using dummy tokens on the source side) can improve MT performance, both the quality and quantity of the backtranslated data play a significant role in practice. Accordingly, the choice of systems to be used for backtranslation is crucial. In Poncelas et al. (2019), different combinations of backtranslated data originating from phrase-based statistical MT (PB-SMT) and neural MT (NMT) were shown to have different impacts on the quality of MT systems.

In this work we conduct a systematic study of the effects of backtranslated data from different sources, as well as how to optimally select subsets of this data taking into account the loss in quality and lexical richness when data is translated with different MT systems. That is, we aim to (i) provide a systematic analysis of backtranslated data from different sources; and (ii) to exploit a reduction in the amount of training data while maintaining high translation quality. To achieve these objectives we analyse backtranslated data from several MT systems and investigate multiple approaches to data selection for backtranslated data based on the Feature Decay Algorithms (FDA: Biçici and Yuret (2015); Poncelas et al. (2018a)) method. We exploit different ways of ranking the data and extracting parallel sentences; we also interleave quality evaluation and lexical diversity/richness information into the ranking process. While our empirical evaluation shows different results for the tested language pairs, this is the first work in this direction and lays a firm foundation for future research.

Nowadays, NMT (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015), and in particular Transformer (Vaswani et al., 2017)

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3898–3908
July 5 - 10, 2020.

achieves state-of-the-art results for many domains and language pairs. However, NMT requires a lot more data than other paradigms (Koehn and Knowles, 2017), which makes it harder to adapt to low-resource scenarios (Sennrich and Zhang, 2019). Using synthetic parallel data via backtranslation has been helpful in some low-resource use-cases (Dowling et al., 2019). For extreme cases with no bilingual parallel corpora, unsupervised MT can obtain reasonable results (Artetxe et al., 2019; Lample and Conneau, 2019). However, its application to real low-resource scenarios is still a matter of study (Marchisio et al., 2020). In this work we are motivated by a real-world low-resource use-case, namely the translation of clinical texts from Basque to Spanish (EU-ES). Basque is a minority language, so most of the Electronic Health Records (EHR) are written in Spanish so that any doctor from the Basque public health service can understand them. The development of a system for translating clinical texts from Basque to Spanish could allow Basque-speaking doctors to write EHRs in Basque, thus contributing to the normalisation of the language in specialised areas.

We conduct our analysis in the scope of the EU-ES translation of EHR use-case, as well as on a language pair and a data set that have been well studied in the literature – German to English (DE-EN) data used in the WMT Biomedical Translation Shared Task (Bawden et al., 2019). As the EU-ES medical data cannot be made publicly available due to privacy regulations, using the DE-EN data is a way to allow for the replicability of our work.

2 Related Work

One of the first papers comparing the performance of different systems for backtranslation was Burlot and Yvon (2018). The authors compared SMT and NMT systems, obtaining similar results. Closer to our work, Soto et al. (2019) also try RBMT, PB-SMT and NMT systems for backtranslating EHRs from Spanish into Basque. However, both papers are limited to comparing the performance of systems trained with backtranslated data originating from a single source, without examining whether a combination might be more effective.

More recently Poncelas et al. (2019) combined the outputs of PB-SMT and NMT systems used for backtranslation, showing that the combination of synthetic data originating from different sources was useful in improving translation performance. In this work we extend these ideas by combining backtranslated data from RBMT, PB-SMT, NMT (LSTM) and NMT (Transformer); in addition, we use FDA to select sentences translated by different systems and analyse the impact of data selection of backtranslated data on the overall translation performance. Regarding the use of data-selection techniques in conjunction with synthetic data, Poncelas and Way (2019) fine-tune NMT models with sentences selected from a backtranslated set, and Chinea-Rios et al. (2017) select monolingual source-side sentences to generate synthetic target strings to improve the translation model.

While the most common approach to assessing the translation capabilities of a MT system is via evaluation scores such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), chrF (Popović, 2015), and METEOR (Banerjee and Lavie, 2005), recently research has begun to address another side of quality of translated text, namely lexical richness and diversity. In a recent paper, Vanmassenhove et al. (2019) study the loss of lexical diversity and richness of the same corpora translated with PB-SMT and NMT systems. Vanmassenhove et al. (2019) investigate the problem for seen (during MT training) and unseen text using MT systems trained on the Europarl corpus (Koehn, 2005), with original (human-produced and translated) text as well as in a round-trip-translation setting.[1] In this work we calculate the same lexical diversity metrics as Vanmassenhove et al. (2019), and further use those metrics to improve the data selection process applied to backtranslated data.

3 Data Selection for Backtranslation from Multiple Sources

FDA (Biçici and Yuret, 2015; Poncelas et al., 2018a) is a data selection technique that retrieves sentences from a corpus based on the number of n-grams overlapping with those present in an in-domain data set referred to as S_{seed}. FDA scores each candidate sentence s according to: (i) the number of n-grams that are shared with the seed S_{seed}; and (ii) the n-grams already present in a set L of

[1] In their experiments, Vanmassenhove et al. (2019) back-translate the training data via an MT system trained on the same data, then train yet another system with this data and analyse its performance. They assess how errors propagate through repeated translation, thereby investigating the extent of inherent algorithm bias in MT models.

selected sentences, as defined in (1):

$$[t]score(s, S_{seed}, L) = \frac{\sum_{ngr \in \{s \bigcap S_{seed}\}} 0.5^{C_L(ngr)}}{\text{length(s)}} \quad (1)$$

where $length(s)$ is the number of words in the sentence s and $C_L(ngr)$ is the number of occurrences of the *n*-gram ngr in L. The score is then used to rank sentences, with the one with the highest score being selected and added to L. This process is repeated iteratively. To avoid selecting sentences containing the same *n*-grams, $score(s, S_{seed}, L)$ applies a penalty to the *n*-grams (up to order three in the default configuration) proportional to the occurrences that have been already selected. In (1), the term $0.5^{C_L(ngr)}$ is used as the penalty.

In the context of MT, FDA has been shown to obtain better results than other methods for data selection (Silva et al., 2018). Acordingly, in this work we too focus on FDA, although our rescoring idea is more general and can be applied to other selection methods based on *n*-gram overlap.

Related work on quality and lexical diversity and richness of MT demonstrates that (i) regardless of the overall performance of an MT system (as measured by both automatic and human evaluation), in general machine-translated text is error-prone and cannot reach human quality (Toral et al., 2018)); and (ii) machine-translated text lacks the lexical richness and diversity of human-translated (or post-edited) text (Vanmassenhove et al., 2019).

In its operation, FDA compares two types of text – the seed and the candidate sentences – without taking into account the quality or the lexical diversity/richness of the candidate text. Our hypothesis is that when selecting data from different sources, FDA cannot account for the differences in quality and lexical diversity/richness of these texts, with the consequence that the selected set (L) is sub-optimal.

We test our hypothesis by assessing the quality and lexical diversity/richness of the backtranslated data with the four different systems as well as with different selected subsets of training data.

To tackle the problem of sub-optimal FDA-selected datasets, we propose to rescore FDA scores based on quality evaluation and lexical diversity/richness scores.[2] That is, for each sentence s_i^{BT} from a backtranslated corpus D_i^{BT} originating from the i^{th} MT system, we factor in the quality expressed by the evaluation metrics, $q(D_i^{BT})$ and the lexical diversity/richness expressed by the diversity metrics, $d(D_i^{BT})$ as shown in (2):

$$score(s_i^{BT}, S_{seed}, L) = \frac{\sum_{ngr \in \{s \bigcap S_{seed}\}} 0.5^{C_L(ngr)}}{\text{length(s)}} \cdot \phi(q(D_i^{BT}), d(D_i^{BT})) \quad (2)$$

where ϕ is a function over quality and lexical diversity metrics producing a non-negative real number.

We note three considerations with respect to our approach to Equation (2).

1. **Sentence-level selection versus document-level quality and lexical diversity/richness evaluation.** The FDA algorithm works on a sentence level, while our approach rescores the FDA scores using document-level metrics. As our goal is to differentiate between the output of different MT systems, we consider metrics that reflect the overall quality of each system. Furthermore, metrics for lexical diversity/richness as type/token ratio (TTR) (Templin, 1975), Yule's I (Yule, 1944), and the measure of textual lexical diversity (MTLD) (McCarthy, 2005) are to be calculated on a document-level; the same is valid for automatic evaluation metrics such as BLEU and TER.
2. **Combined metrics.** We conduct our analysis using the quality metrics BLEU, TER, METEOR and chrF; and TTR, MTLD and Yule's I for lexical diversity/richness. For rescoring we use only BLEU, TER and MTLD as a factor: $\phi = log(BLEU * (100 - TER) * MTLD)$. We decided on this rescoring formula based on preliminary experiments, as it led to the selection of more sentence pairs originating from models trained with backtranslated data from the system that performs best (for both ES-EU and EN-DE); we chose MTLD based on the findings of Vanmassenhove et al. (2019) which show this metric to be more suitable for comparative analysis, as well as mitigating issues related to sentence length typical for TTR and Yule's I (McCarthy, 2005).
3. **Use of devset as a seed**. Using a development set in MT aims to test whether the performance of the MT system has reached a certain level. In

[2] We talk about "rescoring" as if we compare equations (1) and (2), the only difference is the rescoring produced by multiplying equation (1) (left part in equation (2)) by the factors dependent on MT quality and lexical diversity (right part in equation (2)).

FDA for MT, we use a devset as the seed. In our method we compute BLEU and TER on the devset also used as a seed; MTLD is computed on the backtranslated text, i.e. the synthetic source text.

4 Language Pairs – Challenges and Objectives

As a challenging low-resource scenario, we chose the translation of clinical texts from Basque to Spanish, for which there is no in-domain bilingual corpora. We make use of available EHRs in Spanish coming from the hospital of Galdakao-Usansolo to create a synthetic parallel corpus via backtranslation. The Galdakao-Usansolo EHR corpus consists of 142,154 documents compiled between 2008 and 2012. After deduplication, we end up with a total of $2,023,811$ sentences.[3]

As a basis for training the MT systems for backtranslation, we use a bilingual out-of-domain corpus of 4.5M sentence pairs: 2.3M sentence pairs from the news domain (Etchegoyhen et al., 2016), and 2.2M from administrative texts, web-crawling and specialised magazines.

In order to adapt the systems to the clinical domain, we used a bilingual dictionary previously used for automatic clinical term generation in Basque (Perez-de-Viñaspre, 2017), consisting of 151,111 terms in Basque corresponding to 83,360 unique terms in Spanish.

To evaluate our EU-ES systems, we use EHR templates in Basque written with academic purposes (Joanes Etxeberri Saria V. Edizioa, 2014) together with their manual translations into Spanish produced by a bilingual doctor. These 42 templates correspond to diverse specializations, and were written by doctors of the Donostia Hospital. After deduplication, we obtain 1,648 sentence pairs that are randomly divided into 824 sentence pairs for validation (devset) and 824 for testing.

In order to test the generalisability of our idea, we use a well-researched language pair, German-to-English. As our out-of-domain corpus, we used the DE-EN parallel data provided in the WMT 2015 (Bojar et al., 2015) news translation task.

The adaptation of systems to the medical domain with backtranslated data is performed using the UFAL data collection.[4] We selected the following subsets: ECDC, EMEA, EMEA_new_crawl, MuchMore, PatTR_Medical and Subtitles. The total amount of sentences was 2,555,138 which after deduplication was reduced to 2,335,892. After filtering misaligned and empty lines,[5] the resulting amount was 2,322,599 sentences. We used the EN monolingual side. For development and test sets we used the Cochrane and NHS 24 subsets from the Himl 2017 set.[6]

Table 1 provides the statistics of our corpora.

	Desc.	Sent.	Tokens	
			src	trg
EU-ES	out-of-domain	4.5M	73M	102M
	clinical terms	151K	271K	258K
	EHRs	2M		33M
	EHR templates	1.6K	18.5K	17.6K
DE-EN	out-of-domain	4.5M	110M	116M
	in-domain	2.3M		97M
	devset	1K	16K	15K
	test set	467	10K	9.7K

Table 1: Description and statistics of the used corpora.

5 Empirical Evaluation

Via a set of experiments, we (i) investigate the differences in the backtranslated data originating from the four different MT systems and their impact on the performance of MT systems using this backtranslated data, and (ii) test our hypothesis as well as different approaches to rescoring the data selection algorithm.

5.1 Systems Used for Backtranslation

First, we train PB-SMT, LSTM and Transformer models for the ES-EU and EN-DE (i.e. *reverse*) language directions. Then we backtranslate the monolingual corpus into the target language (EU and DE, respectively) using those systems, as well as a RBMT one.

RBMT: We use Apertium (Forcada et al., 2011) for the EN-DE language pair, and Matxin (Mayor, 2007) for ES-EU, adapted to the clinical domain by the inclusion of the same dictionaries used to train the other systems.

PB-SMT: We use Moses with default parameters, using MGIZA for word alignment (Och and Ney,

[3] Due to privacy requirements, this corpus is not publicly available. Prior to use, it was de-identified by reordering sentences, and only authors who had previously signed a non-disclosure commitment had access to it.

[4] https://ufal.mff.cuni.cz/ufal_medical_corpus

[5] We used the clean-corpus-n.pl script provided with the Moses toolkit (Koehn et al., 2007).

[6] http://www.himl.eu/test-sets

2003), an "msd-bidirectional-fe" lexicalised reordering model and a KenLM (Heafield, 2011) 5-gram target language model. We tuned the model using Minimum Error Rate Training (Och, 2003) with an n-best list of length 100.

LSTM: We use an RNN of 4 layers, with LSTM units of size 512, dropout of 0.2 and a batch-size of 128. We use Adam (Kingma and Ba, 2015) as the learning optimiser, with a learning rate of 0.0001 and 2,000 warmup steps.

Transformer: We train a Transformer model with the hyperparameters recommended by OpenNMT,[7] halving the batch-size so that it could fit in 2 GPUs, and accordingly doubling the value for gradient accumulation.

We train all NMT systems using OpenNMT (Klein et al., 2017) for a maximum of 200,000 steps, and select the model that obtains the highest BLEU score on the devset; note that the final systems trained after applying data selection use early stopping with perplexity not decreasing in 3 consecutive steps as our stopping criterion. Backtranslation is performed with the default hyperparameters, including a beam-width of 5 and a batch-size of 30.

We use Moses scripts to tokenise and truecase all the corpora to be used for statistical or neural systems. For the NMT systems, we apply BPE (Sennrich et al., 2016b) on the concatenated bilingual corpora with 90,000 merge operations for EU-ES and 89,500 for DE-EN, using subword-nmt.[8]

5.2 Systems with Data Selected via Backtranslation

For each language pair we train four Transformer models with the authentic and backtranslated data, as well as a fifth system with all four backtranslated versions concatenated to the authentic data. These we refer to as $+S_{bt}$, where S is one of RBMT, PB-SMT, LSTM or Transformer and indicates the origin of the backtranslation, and $+\text{All}_{bt}$ to refer to the system trained with all backtranslated data.

Next, we use the devset as a seed for the data selection algorithm. Given that FDA does not score sentences that have no n-gram overlaps with any sentence from the seed, for the 'EachFromAll' configuration presented later, which is constrained to select one sentence for each sentence in the monolingual corpus, we randomly select one sentence among those produced by the 4 different systems used for backtranslation, in case none of them overlap with any sentence from the seed. We obtain the FDA scores and use them to order the sentence pairs in descending order. Next, we apply the following different data selection configurations:

1. Top from all sentences (referred to as *FromAll* henceforth): concatenate the data backtranslated with all the systems and select the top ranking 2M (for EU-ES) or 2.3M (for DE-EN) sentence pairs with the possibility of selecting the same target sentence more than once, i.e. translated by different systems.
2. Top for each (target) sentence (henceforth, *EachFromAll*): concatenate the data backtranslated with all the systems and select the optimal sentence pairs avoiding the selection of the same target sentence more than once. That is, each selected target sentence will have only one associated source sentence originating from one specific system.
3. Top for each (target) sentence x4 (henceforth, *EachFromAll x4*): same as EachFromAll, but repeating the selected backtranslated data four times (only for EU-ES).
4. Top for each (target) sentence **rescored** (henceforth, *EachFromAll RS*): use MT evaluation and lexical diversity metrics to rescore the FDA ranks and perform an EachFromAll selection.

We selected the Transformer architecture as the basis of our backtranslation models because (i) it has obtained the best performance for many use-cases and language pairs which we also aim at, and (ii) it has been shown that Transformer's performance is strongly impacted by the quantity of data, which can act as an indicator as to whether our improvements originate from the quantity or the quality of the data. That is why we compare EachFromAll systems to systems trained with all backtranslated data (i.e. all 8M sentence pairs), to verify that it is not only the amount of data that impacts performance.

6 Results and Analysis

6.1 MT Evaluation

We use the automatic evaluation metrics BLEU, TER, METEOR and chrF (in its chrF3 variant) to assess the translation quality of our systems. In Table 2 we show the scores on the test set of the

[7] http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model (Accessed on December 9, 2019.)

[8] https://github.com/rsennrich/subword-nmt (Accessed on December 9, 2019.)

reverse systems used for backtranslation (the best are marked in bold). For EU-ES, since we only use clinical terms as in-domain training data, the results are poor overall. However, we observe that Transformer obtains the best results according to all metrics for both EU-ES and DE-EN. Table 3 shows the results of our baseline (*forward*) systems. It shows that Transformer systems perform best for both language pairs. Evaluation scores for the systems trained on authentic and backtranslated data, and for the systems trained after data selection for EU-ES and DE-EN, are shown in Table 4.

		BLEU↑	TER↓	METEOR↑	CHRF3↑
ES-EU	RBMT	11.37	75.52	19.80	41.35
	PB-SMT	9.38	70.70	25.36	44.07
	LSTM	7.01	72.29	20.46	33.94
	Transformer	**12.21**	**66.53**	**26.96**	**44.42**
EN-DE	RBMT	8.21	72.26	25.70	41.40
	PB-SMT	14.85	74.00	35.62	48.92
	LSTM	24.65	54.60	43.30	53.51
	Transformer	**32.24**	**46.83**	**50.25**	**60.29**

Table 2: Scores of *reverse* systems for backtranslation.

		BLEU↑	TER↓	METEOR↑	CHRF3↑
EU-ES	LSTM	10.84	85.00	32.79	41.36
	Transformer	**19.64**	**69.11**	**43.84**	**53.03**
DE-EN	LSTM	28.15	51.95	32.19	55.40
	Transformer	**38.27**	**42.87**	**37.02**	**62.37**

Table 3: Scores of baseline systems.

			BLEU↑	TER↓	MET.↑	CHRF3↑
EU-ES	Auth. + BT.	+RBMT$_{bt}$	23.27	62.67	48.02	56.51
		+PB-SMT$_{bt}$	22.51	64.57	45.97	54.53
		+LSTM$_{bt}$	24.74	63.55	47.58	55.59
		+Transformer$_{bt}$	25.70	60.29	48.53	57.08
		+All$_{bt}$	**26.18**	**59.10**	**49.19**	**57.31**
	Auth.	FromAll	***25.93***	59.76	48.66	56.69
	BT.	EachFromAll	25.85	***58.92***	***48.83***	***57.17***
	+ DS	EachFromAll x4	24.59	61.15	48.10	56.19
		EachFromAll RS	25.77	59.86	48.59	56.92
DE-EN	Auth. + BT	+RBMT$_{bt}$	39.02	42.27	37.32	62.72
		+PB-SMT$_{bt}$	42.32	39.21	39.37	65.91
		+LSTM$_{bt}$	40.97	39.75	38.45	64.81
		+Transformer$_{bt}$	**42.75**	38.73	39.35	**66.05**
		+All$_{bt}$	42.69	**38.45**	**39.65**	65.99
	Auth.	FromAll	43.66	***37.71***	***40.10***	67.01
	+ BT	EachFromAll	43.45	38.24	39.81	66.44
	+ DS	EachFromAll RS	***43.98***	37.79	39.91	***67.10***

Table 4: Scores for systems trained on authentic (Auth.) and backtranslated (BT) data, and after data selection (DS). MET. abbreviates METEOR.

We observe from Table 4 that for both language pairs the inclusion of backtranslated data clearly improves the results of the baseline systems. For EU-ES the ordering of the systems from best to worse is Transformer > RBMT > LSTM > PB-SMT for all metrics except BLEU, where the order is Transformer > LSTM > RBMT > PB-SMT. The EU-ES system trained on (authentic data and) data translated by all systems (+All$_{bt}$), thus using 4 times more backtranslated data than the rest, obtains the best results; however, the observed improvements are not as high as those for the other systems, e.g. the best (+Transformer$_{bt}$) has a 0.96 BLEU point improvement over the second best (+LSTM$_{bt}$), while the +All$_{bt}$ system is only 0.48 BLEU points better than +Transformer$_{bt}$. This tendency is the same for the other metrics too. For the DE-EN use-case the score differences between the best systems (+Transformer$_{bt}$ or +PB-SMT$_{bt}$ depending on the metric) and +All$_{bt}$ are even smaller, with BLEU and chrF3 favouring the former, and TER and METEOR the latter.

For EU-ES, all systems trained with 2M sentence pairs selected from the backtranslated data according to the basic DS methods and the newly proposed method with rescoring obtain better results than any system trained with backtranslated data originating from a *single* system. Furthermore, according to all metrics except BLEU, the EachFromAll system outperforms FromAll. Compared to the system including the data translated by all systems (+All$_{bt}$), EachFromAll is better only in terms of TER. These results show that either the quantity of data leads to differences in performance (comparing the best system after data selection, i.e. EachFromAll, to +All$_{bt}$), or that the data selection method fails to retrieve those sentence pairs that would lead to better performance. In order to test these two assumptions, we first train a system with the EachFromAll data repeated 4 times resulting in the same number of sentence pairs as in the +All$_{bt}$ case. According to the resulting evaluation scores, this system is worse than +All$_{bt}$, but also worse than any of the basic data selection configurations. This indicates that the diversity (among the source sentences) gained by using 4 different systems for backtranslation is more important than the quantity of the data in terms of automatic scores. While for EU-ES the EachFromAll selection configuration achieves the best results, for DE-EN the FromAll configuration leads to better scores. Furthermore, this configuration outperforms the system with all backtranslated data (+All$_{bt}$).

Next, we train a system with data selected from the backtranslated data after the original FDA

scores have been rescored using the quality and lexical diversity/richness scores. These systems are shown in Table 4 with the suffix RS (i.e. ReScored). While for EU-ES this system does not outperform the rest, in the DE-EN case we observe that it does. With the exception of the TER and METEOR scores, the EachFromAll RS for the DE-EN language pair is the best system. These experiments show different outcomes for each language pair and thus disagree with respect to our hypothesis of rescoring the data selection scores being beneficial for MT. Accordingly, more experiments are needed to specify how to perform this rescoring, as well as in which settings our rescoring proposal is beneficial. Further analysis and a discussion on lexical diversity/richness, data selection and sentence length follow in the rest of this section.

6.2 Lexical Diversity/Richness

We analyse the lexical diversity/richness of the corpora of both language pairs based on the Yule's I, MTLD and TTR metrics. We calculate these scores for the corpora resulting from backtranslation by the different systems (BT), for the corpora resulting from applying the basic data selection approaches (DS), and the development and test sets used for evaluation (EV). We show these scores in Table 5 and Table 6 for EU-ES and DE-EN, respectively.

Regarding the different systems used for backtranslation, we observe that for EU-ES the sentences translated by the RBMT system are much more diverse than the rest according to all metrics, while Transformer obtains the highest scores among the other three. For the DE-EN corpora, this is not the case, and the data from the Transformer system is more diverse according to Yule's I and TTR, but not according to MTLD.

We note that Yule's I and TTR depend on the amount of sentences in the assessed corpora. As such, we can see that for the development and test sets the scores are quite a bit higher than the rest. Accordingly, comparisons should be only be conducted for corpora with the same number of sentences.

Following the analysis and discussion in Vanmassenhove et al. (2019), we decided to use MTLD as the lexical diversity metric for our rescoring data selection approach, as defined in Section 3.

6.3 Systems Selected by Data Selection

We first analyse how the basic data selection methods choose different numbers of sentences from each system used for backtranslation, and then we compare them with the rescoring method. Figures 1 and 2 show the portion of selected sentences per backtranslation system that form the training sets for the systems listed in Table 4.

For EU-ES, we observe that the EachFromAll configuration (the one with the highest scores according to the evaluation metrics in Table 4) selects more sentences from Transformer (649,312) in contrast to the ForAll approach that prefers PB-SMT (657,543). For DE-EN, FromAll and EachFromAll tend to select a higher number of sentences backtranslated by the PB-SMT model (820,765 and 924,694, respectively). However, for both language pairs, both ForAll and EachFromAll distributions are very similar as can be seen in Figures 1 and 2. Given that the DE-EN system trained with backtranslated data from PB-SMT (+PB-SMT$_{bt}$) obtains the worst results while the one from Transformer (+Transformer$_{bt}$) performs the best, we correlate the two measurements and hypothesise that a

Type	Corpus	Yule's I*100		MTLD		TTR * 100	
		EU	ES	EU	ES	EU	ES
BT	RBMT$_{bt}$	74.3	0.91	15.33	14.06	3.70	1.01
	PB-SMT$_{bt}$	0.40		13.76		1.01	
	LSTM$_{bt}$	3.23		13.20		2.77	
	Trans.$_{bt}$	8.19		13.79			
DS	FA	2.81	0.16	13.73	13.91	2.26	0.42
	EFA	5.78	0.91	13.88	14.03	3.08	1.01
	EFA RS	9.54	0.91	13.84	14.03	3.67	1.01
EV	Dev.	626	456	13.72	13.92	32.90	27.50
	Test	663	491	13.63	13.75	32.80	27.50

Table 5: Lexical diversity scores of the backtranslation (BT), data selection (DS) and evaluation (EV) corpora for the ES-EU and EU-ES systems. Trans. = Transformer, FA = ForAll, EFA = EachFromAll, EFA RS = EachFromAll Rescored.

Type	Corpus	Yule's I*100		MTLD		TTR * 100	
		DE	EN	DE	EN	DE	EN
BT	RBMT$_{bt}$	4.55	2.68	48.50	37.50	1.64	1.56
	PB-SMT$_{bt}$	0.66		74.90		0.80	
	LSTM$_{bt}$	2.31		40.00		1.90	
	Trans.$_{bt}$	5.62		53.70		2.61	
DS	FA	2.49	0.11	107.00	50	1.44	0.36
	EFA	3.96	0.39	103.00	46.00	1.83	0.69
	EFA RS	5.39	0.39	105.00	45.60	2.56	0.69
EV	Dev	386	282	108.15	61.06	20.00	15.59
	Test	528	301	117.90	59.63	23.83	18.11

Table 6: Lexical diversity scores of the backtranslation (BT), data selection (DS) and evaluation (EV) corpora for the EN-DE and DE-EN systems. Trans. = Transformer, FA = ForAll, EFA = EachFromAll, EFA RS = EachFromAll Rescored.

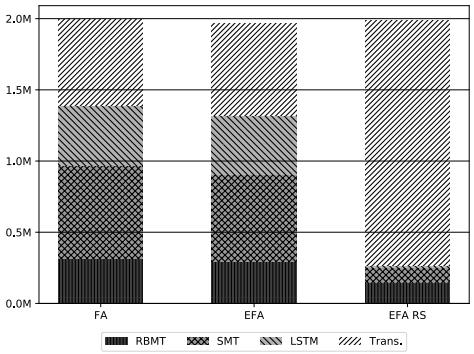

Figure 1: Amount of sentences selected from each system by the data selection approaches for EU-ES. FA = FromAll, EFA = EachFromAll, EFA RS = EachFromAll Rescored.

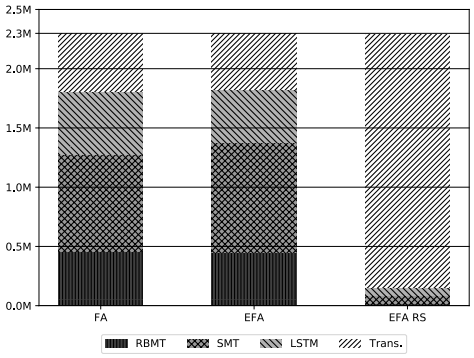

Figure 2: Amount of sentences selected from each system by the data selection approaches for EN-DE. FA = FromAll, EFA = EachFromAll, EFA RS = EachFromAll Rescored.

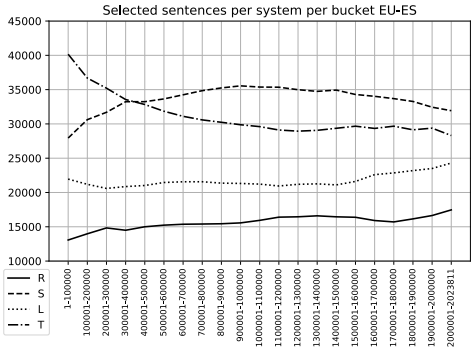

Figure 3: Number of sentences selected from each system by the FromAll data selection approach for EU-ES language pair in subsequent bins of 100,000 sentences (extrapolated for the last bin).

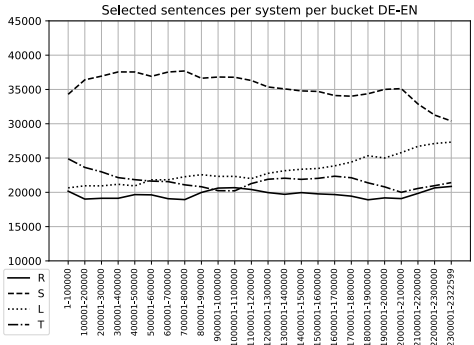

Figure 4: Number of sentences selected from each system by the FromAll data selection approach for DE-EN language pair in subsequent bins of 100,000 sentences (extrapolated for the last bin).

distribution where more sentences originating from Transformer are selected would yield better results. Our ϕ rescoring (cf. Equation (2)) shifts the preferred selection system to Transformer. For EU-ES, the EachFromAll Rescored selects 1,720,736 out of the total of 1,985,227 sentences (about 87%); for DE-EN, it selects 2,131,227 out of the total of 2,284,800 sentences (93%).

For a more in-depth view of the distribution of selected sentence pairs per backtranslation system, we present the amount of selected sentences per system in bins of 100,000 for the FromAll systems. We show the results for EU-ES in Figure 3 and for DE-EN in Figure 4. For EU-ES, we observe that Transformer is the most selected system for the first bins, but the number of sentences sharply decreases until the middle of the corpus and then stabilises. In contrast, the number of sentences originating from PB-SMT increases in the first half and slowly decreases afterwards. The number of sentences from RBMT and LSTM seams more stable, with a slight tendency to increase, peaking in the last bins. For DE-EN, we observe that PB-SMT is always the preferred system, but with a decreasing tendency; and the number of sentences originating from LSTM increases towards the last bins.

6.4 Sentence Length

We also analyse how the average sentence length varies during the data selection process in the FromAll configuration, as we did in Section 6.3 when analysing the selected systems.

Table 7 shows the average sentence lengths of the EU-ES and DE-EN data from the different reverse systems (BT), of the corpora resulting after data selection (DS) and of the test and the development sets (EV). We note that the sentences translated by PB-SMT are longer than those translated

by any other system for both language pairs. Correlating these results with those presented in Table 4 and in Figures 3 and 4, we can assert that in FDA the length penalty has a weaker effect than n-gram overlap and as such FDA has a preference towards n-gram MT paradigms, i.e. PB-SMT. However, data selection that results in more Transformer sentences would appear to be a better option.

Type	Corpus	EU	ES	DE	EN
BT	RBMT$_{bt}$	10.56	16.16	33.64	34.30
	PB-SMT$_{bt}$	16.09	16.16	39.04	34.30
	LSTM$_{bt}$	12.53	16.16	29.55	34.30
	Transformer$_{bt}$	12.62	16.16	23.37	34.30
DS	FromAll	17.60	21.21	41.61	51.84
	EachFromAll	13.67	16.16	32.94	34.30
EV	Dev.	10.85	10.34	15.09	14.34
	Test	11.64	11.04	21.27	20.79

Table 7: Average sentence length of the backtranslation (BT), data selection (DS) and evaluation sets (EV).

7 Conclusions and Future Work

We evaluated several approaches to data selection over the data backtranslated by RBMT, PB-SMT, LSTM and Transformer systems for two language pairs (EU-ES and DE-EN) from the clinical/biomedical domain. The former is a low-resource language pair, and the latter a well researched, high-resource language pair. Furthermore, in terms of the two target languages, English is a morphologically less rich language than Spanish, which creates a different setting again in which to evaluate our methodology. We use these two different use-cases to better understand both data selection and backtranslation.

We show how the different FDA data selection configurations tend to select different numbers of sentences coming from different systems, resulting in MT systems with different performance.

Under the assumption that FDA's performance is hindered by the fact that the data originates from MT systems, and as such contains errors and is of lower lexical richness, we rescored the data selection scores for each sentence by a factor depending on the BLEU, TER and MTLD values of the system used to backtranslate it. By doing so, we managed to improve the results for the DE-EN system, while for EU-ES we obtained similar performance to the other MT systems; this allows us to use just 25% of the data. Further investigation is required to study under which conditions our proposed rescoring method is beneficial, but our experiments with both low- and high-resource language pairs suggest that if the systems used for backtranslation are poor, then this technique will be of little value; clearly this is closely related to the amount of resources available for the language pair under study.

In the future, we plan to investigate ways to directly incorporate the rescoring metrics into the data selection process itself, so that penalising similar sentences can also be taken into account. We also aim to conduct a human evaluation of the translated sentences in order to obtain a better understanding of the effects of data selection and backtranslation on the overall quality. Finally, we intend to analyse the effect of these measures in a wider range of language pairs and settings, in order to propose a more general solution.

Acknowledgements

Xabier Soto's work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) FPI grant number BES-2017-081045. This work was mostly done during an internship at the ADAPT Centre in DCU.

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA. 15pp.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the

WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy.

Ergun Biçici and Deniz Yuret. 2015. Optimizing instance selection for statistical machine translation with feature decay algorithms. *Transactions on Audio, Speech & Language Processing*, 23(2):339–350.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal.

Franck Burlot and François Yvon. 2018. Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Belgium, Brussels.

Mara Chinea-Rios, Alvaro Peris, and Francisco Casacuberta. 2017. Adapting neural machine translation with parallel synthetic data. In *Proceedings of the Second Conference on Machine Translation*, pages 138–147, Copenhagen, Denmark.

Meghan Dowling, Teresa Lynn, and Andy Way. 2019. Investigating backtranslation for the improvement of English-Irish machine translation. *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 26:1–25.

Thierry Etchegoyhen, 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)*, pages 3523–3529, Portoroz, Slovenia.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: A free/open-source platform for rule-based machine translation. *Neural Computation*, 25(2):127–144.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, UK.

Chris Hokamp. 2017. Ensembling factored neural machine translation models for automatic post-editing and quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 647–654, Copenhagen, Denmark.

Joanes Etxeberri Saria V. Edizioa. 2014. Donostia unibertsitate ospitaleko alta-txostenak. *Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, , Washington, USA.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA. 15pp.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: The tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Computing Research Repository*, arXiv:1901.07291.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *Computing Research Repository*, arXiv:2004.05516.

Aingeru Mayor. 2007. *Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz*. Ph.D. thesis, University of the Basque Country, Donostia, Spain.

Philip M McCarthy. 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity*. Ph.D. thesis, University of Memphis, TN.

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Olatz Perez-de-Viñaspre. 2017. *Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque*. Ph.D. thesis, University of the Basque Country, Donostia, Spain.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2018a. Feature decay algorithms for neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 239–248, Alicante, Spain.

Alberto Poncelas, Maja Popovic, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Combining SMT and NMT Back-Translated Data for Efficient NMT. In *Proceedings of Recent Advances in Natural Language Processing*, pages 922–931, Varna, Bulgaria.

Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018b. Investigating backtranslation in neural machine translation. In *21st Annual Conference of the European Association for Machine Translation*, pages 249–258, Alicante, Spain.

Alberto Poncelas and Andy Way. 2019. Selecting Artificially-Generated Sentences for Fine-Tuning Neural Machine Translation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 219–228, Tokyo, Japan.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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, Berlin, Germany.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy.

Catarina Cruz Silva, Chao-Hong Liu, Alberto Poncelas, and Andy Way. 2018. Extracting in-domain training corpora for neural machine translation using data selection methods. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 224–231, Brussels, Belgium.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, USA.

Xabier Soto, Olatz Perez-De-Viñaspre, Maite Oronoz, and Gorka Labaka. 2019. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 8–18, Dublin, Ireland.

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

Mildred C. Templin. 1975. *Certain Language Skills in Children: Their Development and Interrelationships*. University of Minnesota Press, Minneapolis, MN.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Belgium, Brussels.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII (Research Track)*, pages 222–232, Dublin, Ireland.

Ashish Vaswani, 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*, pages 5998–6008, Long Beach, CA.

Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 101–105, Florence, Italy.

G. Udny Yule. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, UK.

Ixamed's submission description for WMT20 Biomedical shared task: benefits and limitations of using terminologies for domain adaptation

Xabier Soto, Olatz Perez-de-Viñaspre, Gorka Labaka, Maite Oronoz
HiTZ Basque Center for Language Technologies - Ixa, University of the Basque Country UPV/EHU
{xabier.soto, olatz.perezdevinaspre, gorka.labaka, maite.oronoz }@ehu.eus

Abstract

In this paper we describe the systems developed at Ixa for our participation in WMT20 Biomedical shared task in three language pairs, en-eu, en-es and es-en. When defining our approach, we have put the focus on making an efficient use of corpora recently compiled for training Machine Translation (MT) systems to translate Covid-19 related text, as well as reusing previously compiled corpora and developed systems for biomedical or clinical domain. Regarding the techniques used, we base on the findings from our previous works for translating clinical texts into Basque, making use of clinical terminology for adapting the MT systems to the clinical domain. However, after manually inspecting some of the outputs generated by our systems, for most of the submissions we end up using the system trained only with the basic corpus, since the systems including the clinical terminologies generated outputs shorter in length than the corresponding references. Thus, we present simple baselines for translating abstracts between English and Spanish (en/es); while for translating abstracts and terms from English into Basque (en-eu), we concatenate the best en-es system for each kind of text with our es-eu system. We present automatic evaluation results in terms of BLEU scores, and analyse the effect of including clinical terminology on the average sentence length of the generated outputs. Following the recent recommendations for a responsible use of GPUs for NLP research, we include an estimation of the generated CO_2 emissions, based on the power consumed for training the MT systems.

1 Introduction

The WMT20 Biomedical shared task calls for developing systems for translating biomedical abstracts and terminologies between several languages. In our case, we participate in the task of translating biomedical terms and abstracts from English into Basque (en-eu), as well as translating biomedical abstracts between English and Spanish (en-es and es-en). For translating the test data from English into Basque, we concatenate our best en-es system with our es-eu system, both for translating abstracts and terminologies.

2 Related work

For translating biomedical texts from English into Catalan, Costa-jussá et al. (2018) use a pivoting or cascade approach, translating the texts first from English into Spanish (en-es), and then from Spanish into Catalan (es-ca). This technique is useful when there are more bilingual in-domain sentences for each of the language pairs (en/es and es/ca) than for the desired source and target languages (en/ca). Since there are low resources for en/eu biomedical domain, but we have access to many resources for en/es and es/eu in the biomedical or clinical domain, we follow the same approach for translating the test sets from English into Basque (en-eu).

Since most of the available in-domain corpus is monolingual, we also make use of traditional back-translation and forward translation techniques (Sennrich et al., 2016).

In our previous work for translating clinical texts between Basque and Spanish, we showed that including clinical terminologies directly into the training corpus was useful for domain adaptation when no bilingual in-domain sentences were available (Soto et al., 2019a). As clinical terminologies, we refer to the automatic translation into Basque of SNOMED CT (IHTSDO, 2014), which is considered the most comprehensive, multilingual clinical health care terminology collection in the world. In this work, we extend the number of clinical terminologies as part of the ongoing translation of SNOMED CT into Basque (Perez-de-Viñaspre,

Proceedings of the 5th Conference on Machine Translation (WMT), pages 875–880

2017), and include the provided ICD-10 resources plus other smaller terminology collections recently created for translating Covid-19 related texts.

3 Resources

For training our baseline en/es systems, we make use of the Medline corpus provided by the organisers of the WMT20 Biomedical shared task, as well as the recently compiled TAUS Corona Crisis Corpus.[1]

For backtranslation (es-en) and forward translation (en-es), we use the English corpus prepared by Sketch Engine[2], based on the Covid-19 related corpus compiled for a recent Kaggle competition (Wang et al., 2020).

As a final step, we include several clinical terminologies: 1) the ICD-10 (en-eu) corpus provided by the organisers of the WMT20 Biomedical shared task, adding the corresponding Spanish counterparts; 2) terms obtained from the automatic translation into Basque of SNOMED CT (Perez-de-Viñaspre, 2017), including terms up to 11 tokens; 3) a recent SNOMED CT interim release of Covid-19 related terms[3], manually translated into Basque by a translator of the Basque public health service (Osakidetza); and 4) a collection of Covid-19 related terms recently compiled by Elhuyar[4], including all the terms published until June 18[5].

For training our es-eu system, we use the aforementioned terminologies together with an out-of-domain corpus formed mainly by news (Etchegoyhen et al., 2016), previously applying a language identification tool[6] to exclude sentences where most of the terms are named entities like locations or person names. Doing this, a bigger part of the vocabulary can be used to translate biomedical or clinical terms. Furthermore, as in-domain corpus we use clinical notes in Spanish coming from the hospital of Galdakao-Usansolo for forward translation and copying (Currey et al., 2017). This corpus was compiled between 2008 and 2012.[7]

For the evaluation of en/es systems, we use Khresmoi;[8] while for es-eu we use templates of clinical notes in Basque written in the Donostia hospital (Joanes Etxeberri Saria V. Edizioa, 2014), together with their manual translations into Spanish made by a bilingual doctor.

Table 1 presents the description and statistics of our corpora.

	Description	Sentences
en/es	Medline (WMT Biomedical)	388,068
	TAUS Corona Crisis Corpus	902,133
	Sketch Engine Covid-19 (en)	4,671,609
	ICD-10 (WMT Biomedical)	27,696
	SNOMED CT corpus	385,800
	SNOMED CT Covid-19 corpus	84
	Elhuyar Covid-19 corpus	113
	Khresmoi (dev set)	500
	Khresmoi (test set)	1,000
es-eu	out-of-domain	3,703,757
	in-domain (es)	2,023,811
	ICD-10 (WMT Biomedical)	27,696
	SNOMED CT corpus	896,898
	SNOMED CT Covid-19 corpus	84
	Elhuyar Covid-19 corpus	126
	Donostia hospital (dev set)	1,038
	Donostia hospital (test set)	1,038

Table 1: Description and statistics of the used corpora.

4 Systems

For en/es we develop 3 systems: 1) using only the bilingual in-domain corpus (Medline + TAUS Corona Crisis Corpus), 2) including the Sketch Engine Covid-19 (en) corpus for backtranslation (es-en) or forward translation (en-es), and 3) adding all the clinical terminologies from ICD-10, SNOMED CT and Elhuyar.

For es-eu we train a unique system using the out-of-domain corpus and the clinical terminologies, as well as the in-domain (es) corpus for forward translation and copying.

For training the backtranslation (en-es) and forward translation (es-en) systems, we used the bilingual in-domain corpus (Medline + TAUS Corona Crisis Corpus); while for es-eu we used the out-of-domain corpus and a reduced set of SNOMED CT terminologies, as used in Soto et al. (2019b).

[1] https://md.taus.net/corona

[2] https://www.sketchengine.eu/covid19/

[3] http://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release%2DCOVID-19

[4] https://www.elhuyar.eus/eu/site/prentsa-aretoa/368/covid-19-gaitzaren-inguruko-terminologia%2Dgure-hiztegietako-azkenaldaketak

[5] when the English term was missing, if there was no doubt about how to translate it, the first author manually translated it; while if there wasn't a clear translation into English or the term was more related to socioeconomics than biomedical domain, it wasn't included in the en/es corpus.

[6] https://github.com/saffsd/langid.py

[7] Due to privacy requirements, this corpus is not publicly available. Prior to use, it was de-identified by reordering sentences, and only authors who had previously signed a non-disclosure commitment had access to it.

[8] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122

All the systems are Transformer (Vaswani et al., 2017) models trained with OpenNMT (Klein et al., 2017), using the recommended hyperparameters.[9] When necessary, we halved the batch-size so that it could fit in 2 GPUs, and accordingly doubled the value for gradient accumulation.

We applied joint BPE-dropout (Provilkov et al., 2020), with 32,000 merge operations for en/es and 90,000 for es-eu.

5 Results

Table 2 shows the BLEU scores of our systems on the validation (dev) and test sets presented in Table 1, together with previously published (es-eu) results for comparison.

Lang.	System	dev	test
es-en	Baseline (Medline + TAUS)	56.57	52.55
	Baseline + backtranslation (bt)	**61.60**	**57.25**
	Baseline + bt + terminologies	60.95	56.89
en-es	Baseline (Medline + TAUS)	48.02	46.30
	Baseline + forward translation (ft)	**50.20**	**47.19**
	Baseline + ft + terminologies	49.92	47.15
es-eu	Soto et al. (2019a)	11.30	**12.04**
	Soto et al. (2019b)	**11.85**	11.24
	This work	6.21	5.15

Table 2: BLEU scores for systems developed for es-en, en-es and es-eu translation directions (Lang.).

As expected, backtranslation significantly improves the es-en results (around 5 BLEU points); while the gains obtained with forward translation (en-es) are smaller (around 2 BLEU points in the dev set and around 1 BLEU point in the test set). However, we observe a slight decrease on BLEU values when including the clinical terminologies on the training corpus for both es-en and en-es systems. For further analysing this, we calculate the average sentence length of the different evaluation corpora as translated by the different systems. Table 3 shows the average sentence length of the validation (dev) and test sets after being translated by each of the es-en and en-es systems. As a reference, the average sentence length of the original dev and test sets are 22.70 (es) / 21.06 (en) and 24.03 (es) / 21.91 (en).

We observe that, except for the dev set translated by the en-es systems, the lower sentence length is always obtained when using the system including the clinical terminologies. This is confirmed by a fast check of the outputs generated when translating

[9] http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model (Accessed on July 18, 2020.)

Lang.	System	dev	test
es-en	Baseline (Medline + TAUS)	20.54	22.02
	Baseline + backtranslation (bt)	20.56	21.73
	Baseline + bt + terminologies	20.40	21.56
en-es	Baseline (Medline + TAUS)	22.75	23.87
	Baseline + forward translation (ft)	22.93	23.84
	Baseline + ft + terminologies	22.99	23.76

Table 3: Average sentence length of the different evaluation corpora as translated by the systems developed for es-en and en-es translation directions (Lang.).

the official test sets provided by the organisers, where we see that the sentences translated by these systems usually end before having translated all of the terms that appear in the input. Overall, the sentence lengths of the generated translations are closer to the original sentence lengths when using the baseline systems; therefore, for en-es and es-en we submit as best systems the translations produced by the baseline systems, using only Medline and TAUS corpora.

Regarding es-eu, in Table 2 we can see a severe decrease on BLEU scores comparing to our previous works. For training the system in Soto et al. (2019a) we used the same out-of-domain corpus (without applying langid.py) and a reduced set of SNOMED CT terminologies (151,111 entries), both directly and inserted into artificial sentences; while in Soto et al. (2019b) we used this same corpus without the artificial sentences, which didn't prove to be useful. Nevertheless, after manually checking the outputs generated by these 3 systems, we observe that the system developed for this work performs generally better, so we submit the translations produced by this system.[10] As we use a cascade approach for en-eu, we use the en-es system including the terminologies for translating abstracts; and the baseline system for translating terminologies, as these were the best performing systems on a fast human evaluation. [11]

Once we have selected the best performing systems for each of the language pairs, since we are allowed to submit 3 runs, in the case of en/es, for each of the developed systems we submit an ensemble of the 3 models which obtained higher BLEU

[10] It has to be noted that the evaluation corpus used for es-eu has strong limitations, since the original sentences are written for encouraging medicine students to write correctly; while the translations into Basque made by a doctor are overall shorter, use simplified grammar, often omit verbs and punctuation, and use many acronyms.

[11] Both for en/es and en-eu systems, the translations of the first 10 sentences of the official test sets were checked; and in case of tie, the next 10 sentences were also observed.

scores in the dev set during training; while for en-eu we alternate between single and ensemble systems for each of the en-es and es-eu systems. Specifically, we submit as best system an ensemble of the baseline en-es system and a single es-eu system for translating terminologies; while we use a single en-es system including the terminologies and an ensemble es-eu system for translating abstracts.

Table 4 shows the BLEU scores obtained on the official test sets for each of the language pairs and submitted runs for translating abstracts, as provided by the organisers. We present in italics the result of the expected best system for each language pair, and in bold the highest BLEU score, as in previous tables.

Lang.	System	BLEU
es-en	Baseline (Medline + TAUS)	*40.65*
	Baseline + backtranslation (bt)	**40.71**
	Baseline + bt + terminologies	39.96
en-es	Baseline (Medline + TAUS)	***41.71***
	Baseline + forward translation (ft)	38.36
	Baseline + ft + terminologies	38.58
en-eu	single (en-es) + ensemble (es-eu)	*8.15*
	ensemble (en-es) + single (es-eu)	7.82
	ensemble (en-es) + ensemble (es-eu)	**8.84**

Table 4: BLEU scores on the official test sets for translating abstracts in es-en, en-es and en-eu translation directions (Lang.).

Comparing to the submissions made by other teams, our systems submitted for en/es obtain the lowest BLEU scores among all the participants; while for en-eu our best run is the second among the best runs of each participant, only surpassed by the three runs submitted by Elhuyar.

Finally, Table 5 presents the accuracy and BLEU scores obtained by our systems when used for translating terminologies (en-eu), as provided by the organisers.

Lang.	System	Acc.	BLEU
en-eu	single (en-es) + ensemble (es-eu)	0.12	13.14
	ensemble (en-es) + single (es-eu)	*0.08*	*7.21*
	ensemble (en-es) + ensemble (es-eu)	**0.13**	**14.81**

Table 5: Accuracy (Acc.) and BLEU scores on the official test set for translating terminologies in en-eu translation direction (Lang.).

Surprisingly, the obtained automatic scores are much lower than the ones obtained by the rest of the participants (between 0.73 and 0.78 for accuracy, and approximately 71 to 74 BLEU scores). However, the generated translations look quite sensible, so we expect the human evaluation will shed some light about the performance of our systems.

6 Measured power consumption and estimated CO_2 emissions

Following the recommendations by Strubell et al. (2019), we report the power consumed by our GPUs when training the systems developed for this work, along with the estimated CO_2 emissions. For calculating the training time, we use the time shown in the first and last lines of the log file generated while training the systems, including also the initial time for preparing the data, so the presented values constitute an upper bound of the actually consumed power. Nonetheless, we have to point out that OpenNMT makes an efficient use of the power capabilities of the GPUs, so we can say that the numbers shown here are an accurate estimation. Table 6 shows the number of GPUs, training time, power consumption and estimated CO_2 emissions for each of the developed systems. All the GPUs used for this work are Nvidia Titan Xp models with 250W power. We present the values of the different systems in the same order as in Table 2, and estimate the CO_2 emissions by applying equations (1) and (2) in Strubell et al. (2019), considering only the power consumed by our GPUs. Overall, the CO_2 emissions generated by our GPUs are approximately 329.44 lbs.

Lang.	GPUs	Time (hh:mm)	Power (kWh)	CO_2e (lbs)
es-en	4	43:19	43.33	65.31
	2	46:30	23.26	35.06
	2	45:37	22.82	34.39
en-es	4	45:09	45.16	68.07
	2	47:24	23.70	35.73
	2	47:21	23.68	35.69
es-eu	2	73:14	36.62	55.20
TOTAL				**329.44**

Table 6: Number of GPUs, training time, power consumption and estimated CO_2 emissions for each of the developed systems (same order as in Table 2).

7 Conclusion and future work

In this work, we have presented a simple proposal using previously compiled corpora from the biomedical or clinical domain, as well as clinical terminology included directly to the training corpora. Apart from calculating BLEU scores, we have also calculated the average sentence length of the generated translations for en/es systems, and observed that the systems including terminologies

performed generally worse than the baseline systems.

As future work, we plan to incorporate these clinical terminologies in a more efficient way (Dinu et al., 2019; Wang et al., 2019). For improving both training and evaluation, we'll also use bilingual clinical domain corpora being compiled now in collaboration with the Basque public health service (Osakidetza). Furthermore, since we have observed that some of the translations generated by the es-eu systems remain in Spanish, we'll study techniques to leverage in-domain monolingual data in Basque like the one provided by the organisers from Wikipedia.

Finally, we plan to keep reporting the consumed power and consequently generated CO_2 emissions, probably making use of recently developed automatic tools (Henderson et al., 2020)[12].

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) FPI grant number BES-2017-081045, and projects BigKnowledge (BBVA foundation grant 2018), DOMINO (PGC2018-102041-B-I00, MCIU/AEI/FEDER, UE) and DOTT-HEALTH (PID2019-106942RB-C31, MCIU/AEI/FEDER, UE).

References

Marta R. Costa-jussá, Noé Casas, and Maite Melero. 2018. English-catalan neural machine translation in the biomedical domain through the cascade approach. *Computing Research Repository*, arXiv:1803.07139. Version 2.

Anna Currey, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Thierry Etchegoyhen, 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)*, pages 3523–3529, Portoroz, Slovenia.

Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Computing Research Repository*, arXiv:2002.05651.

International Health Terminology Standards Development Organisation IHTSDO. 2014. *SNOMED CT Starter Guide*. Technical report, International Health Terminology Standards Development Organisation.

Joanes Etxeberri Saria V. Edizioa. 2014. Donostia unibertsitate ospitaleko alta-txostenak. *Donostiako Unibertsitate Ospitalea, Komunikazio Unitatea*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72, Vancouver, Canada.

Olatz Perez-de-Viñaspre. 2017. *Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque*. Ph.D. thesis, University of the Basque Country, Donostia, Spain.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. 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. Association for Computational Linguistics.

Xabier Soto, Olatz Perez-De-Viñaspre, Gorka Labaka, and Maite Oronoz. 2019a. Neural machine translation of clinical texts between long distance languages. *Journal of the American Medical Informatics Association*, 26(12):1478–1487.

Xabier Soto, Olatz Perez-De-Viñaspre, Maite Oronoz, and Gorka Labaka. 2019b. Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 8–18, Dublin, Ireland.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *Computing Research Repository*, arXiv:1906.02243.

[12] https://github.com/Breakend/experiment-impact-tracker

Ashish Vaswani, 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*, pages 5998–6008, Long Beach, CA.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset. *Computing Research Repository*, arXiv:2004.10706. Version 4.

Tao Wang, Shaohui Kuang, Deyi Xiong, and António Branco. 2019. Merging external bilingual pairs into neural machine translation. *Computing Research Repository*, arXiv:1912.00567.

Chapter X: Development of a Machine Translation system for promoting the use of a low resource language in the clinical domain: the case of Basque

Author(s):

*Xabier Soto (HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU, xabier.soto@ehu.eus)

Olatz Perez-de-Viñaspre (HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU, olatz.perezdevinaspre@ehu.eus)

Maite Oronoz (HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU, maite.oronoz@ehu.eus)

Gorka Labaka (HiTZ Basque Center for Language Technologies - Ixa NLP Group, University of the Basque Country UPV/EHU, gorka.labaka@ehu.eus)

Abstract

In multilingual environments where there is a strong language spoken by a majority and a low resource language spoken by a minority, Machine Translation (MT) can be useful for allowing clinical practitioners to write their reports in the minority language, which then can be automatically translated into the majority language.

Current state-of-the-art approaches for MT require large quantities of parallel sentences of the desired languages and the specific domain, so MT systems developed for translating clinical text from/into a low resource language will usually need to go through a domain adaptation process. When there are enough in-domain resources for the target language, back-translation is commonly used for domain adaptation. In our case, since we have access to many Electronic Health Records (EHR) in Spanish, we make use of this and similar techniques that leverage monolingual data for translating clinical text from Basque into Spanish.

Moreover, one of the main characteristics of clinical domain text is its rich terminology, which is not always available for any given language, so before developing an MT system for the clinical domain, it is beneficial to make a special effort to translate the clinical terminology into the low resource language.

If the final objective is to implement a system that can be useful for clinical practitioners, it is important to work with them for defining the terminology and any other aspect that can affect the final performance of the systems. Needless to say, given the special relevance of the content to be translated, users of the MT systems should be aware of possible errors made; and whenever it is possible, a human translator should review the generated translations to guarantee their accuracy.

In this Chapter we describe the approach we have followed to develop an MT system for translating clinical text from Basque into Spanish. In the first section we introduce Basque language and give some details about the sociolinguistic situation in the Basque Country. In the second section we present Itzulbide, the project carried out together with the Basque public health service for compiling clinical domain corpora to be used for developing an MT tool for the healthcare domain. In the third section we overview the diverse corpora we have used in our systems, and specify how the training and evaluation are performed. The fourth section discusses the results obtained in the defined settings. Finally, the fifth section presents some conclusions and points to possible future directions.

1 Introduction

Basque is a pre-Indo-European language spoken in the Basque Country, a region spanning an area in northeastern Spain and southwestern France, with a population of around 3M people[1]. Nowadays, Basque language is spoken by 28,4% of the people in all territories (751,500 active speakers and 1,185,500 passive). Of these active speakers, 700,300 live in the Spanish part (Basque Country and Navarre autonomous communities) and the remaining 51,200 live in the French part (‘Euskal Hirigune Elkargoa’ in Basque, or ‘Communauté d'agglomération du Pays Basque’ in French)[2].

We can then state that Basque is a minority language that persists surrounded by two powerful languages, Spanish and French. Linguistically, Basque is considered an isolated language of unknown origin, and so, it does not share any characteristic with its neighboring Romance languages. In the following examples the reader may recognize the distance among these languages, along with the English glosses for reference:

Basque

buru-ko	*mina*	*dauka*	*eta*	*hiru*	*egun*	*daramatza*	*botaka*	.
head-GEN	pain	has	and	three	days	carries	vomiting	.

‘he/she has a headache and has been vomiting for three days.’

Spanish

le	*duele*	*la*	*cabeza*	*y*	*lleva*	*tres*	*días*	*vomitando*	.
to-him/her	hurts	the.F	head	and	carries	three	days	vomiting	.

1 https://en.wikipedia.org/wiki/Basque_Country_(greater_region)
2 https://en.wikipedia.org/wiki/Basque_language

‘he/she has a headache and has been vomiting for three days.’

French

il/elle	*a*	*un*	*mal-de-tête*	*et*	*vomit*	*depuis*	*trois*	*jours*	.
he/she	has	a	headache	and	vomits	since	three	days	.

‘he/she has a headache and has been vomiting for three days.’

Nowadays, Basque holds co-official language status with Spanish in the Basque Autonomous Community and in the northern part of Navarre, but during centuries Basque has not been an official language; it was out of educational systems, out of media, and out of industrial environments, being its use mostly limited to the private sphere. It was not until the late 1960s that the standardization process of Basque language started, while the current autonomous institutions, including the health related ones, were created in the early 1980s. Due to these and other features, the use of Basque language in the bio-sanitary system is still low.

In the Basque Autonomous Community, the oral communication between patients and healthcare workers (specially nurses and physicians) can be done in Basque given that both sides are Basque speakers, but almost all of the reports are still written in Spanish. This is often caused by a lack of habit of the healthcare workers to write in Basque, owing to diverse reasons like the language used in education, knowledge of health related terminology, etc. However, even when the healthcare workers want to write the reports in Basque, the diglossic situation, where everyone knows Spanish and only a minority knows Basque, makes them

feel forced to write the documents in Spanish. One of the reasons for this is that Osakidetza (the public health service in the Basque Autonomous Community) has a centralized system for storing the health information of the patients, so healthcare workers can access the clinical records of the patients, which have been possibly written by another colleague. Thus, when a healthcare worker wants to write a report in Basque, if the following readers are not Basque speakers, the safety of the patient can be put at risk.

For comparison[3], we have studied how the communication between patients and healthcare workers is done in other multilingual countries. In Canada, in the areas where more than one language is official, patients decide the language they want to use in their communications (Desjardins 2003), so all their health records are written in that language. By contrast, in Belgium, the communities are separated by the language they use, each of them having their own public health service. In the case of Brussels, being a bilingual area, health services are offered both in French and Flemish (Gerkens and Merkur 2010). As a last example, in Luxembourg, where German, French, Italian, English and Portuguese are all widely used languages, they use French as lingua franca for communications in the health care context (European Observatory on Health Care Systems 1999). In the Basque Country, the language communities are merged, using Spanish as lingua franca in the southern part of the Basque Country and French in the northern part. In this situation, the linguistic rights of the Basque speaking patients and healthcare workers are not preserved, since, as said before, even if the oral communication can be done in Basque, the whole clinical attention can not be given in Basque. This work summarises the first steps done for creating the necessary conditions for

3 This paragraph has been adapted from Perez-de-Viñaspre (2017)

Basque speaking healthcare workers to write their reports in Basque, while guaranteeing that the safety of the patients is not put at risk.

2 Itzulbide

Itzulbide is a project defined in cooperation between Osakidetza and the University of the Basque Country that aims to promote the use of Basque in the EHRs. As mentioned before, since the EHR storage system in Osakidetza is centralized, any next doctor to see the patient may access all the previous EHRs. In the case of monolingual speakers of Spanish, they may not understand what is written in Basque. This project wants to take steps to address this problem, creating an MT system from Basque to Spanish for the clinical domain.

In order to train the automatic translator, it is essential to collect parallel medical reports in Basque and Spanish. To this end, work has been done on a web application that helps healthcare workers to write bilingual EHRs.

In subsection 2.1 we make a general description of the Itzulbide project, in subsection 2.2 we specify how the compiled documents are classified, and in subsection 2.3 we briefly present the web application designed for collecting the bilingual corpus to be used by the MT system.

2.1 Project description

One of the main challenges of Osakidetza is to progressively increase the presence of Basque in clinical documentation and medical records. To this end, Osakidetza proposed a public bid which the HiTZ group won and later became the Itzulbide project.

There is a growing demand from Osakidetza's healthcare workers for the use of the Basque language in medical records, but this does not materialise as they are aware that a large part of the professionals do not speak both languages. In addition, the progressive increase in the use of Basque in oral communications between professionals and between professional and patient leads to a natural increase in the use of Basque in the written section. However, in practice, it happens that it requires an additional effort on the part of the professional, who in the context of a consultation that is being carried out in Basque, has to transcribe its content in Spanish in order to guarantee the continuity of the patient's care. This, in turn, does not encourage the professional to use the patient's preferred language.

The moment to face this project in 2018 was considered adequate for two main reasons:

1. 46.60% of active employees in structural positions and 56.29% of the temporary staff were bilingual at that moment.
2. the paradigm shift that has involved the use of neural networks and deep learning in Natural Language Processing (NLP) in general, and in machine translation in particular, changed, obtaining very high quality translations.

Although there has been an accelerated growth in the development of machine translation systems, these do not cover the specific needs that may arise in the field of healthcare. There are no tools that respond to the Osakidetza healthcare context with the required reliability, accuracy and quality.

Taking into account all these facts, the development of a specific MT tool was considered essential to maximize the accuracy and reliability of text translated from Basque into Spanish based on a bilingual corpus of clinical texts.

Through this project, we contribute to Osakidetza, not only by responding to the strategic challenges set out in its linguistic policy, but also facing a growing social demand. The quality of care is increased by improving the communication between professionals and patient-professionals and enhancing their satisfaction. The cost/benefit ratio is also improved as external translation costs are reduced.

The project is divided in three phases:

1. Corpus creation: volunteer health care workers are recruited to write bilingual EHRs (Basque/Spanish).
2. MT training: using the bilingual corpus and other available bilingual and monolingual resources, the neural machine translator is trained.
3. MT evaluation: part of the volunteers manually evaluate the system.

2.2 Classification of the medical records

Classifying the documents considering different aspects will help us to perform diverse experiments when training and evaluating the neural translator. The web application for compiling the bilingual corpus considers the characteristics described in the following lines. Volunteers participating in the project will need to create an account in the web application designed for the corpus collection. When logging in for the first time, each healthcare worker

will indicate the health organization in which they work. Osakidetza is organized in 'integrated health organizations', known by their acronym in Spanish OSI ('*Organización de Servicio Integrado*'), that aim to integrate the different levels of patient care, and group hospitals and health centers in the same organization. They are organized by geographical areas. For example the OSI in the region around Donostia is composed of 32 health centers and outpatient clinics, and a university hospital. The scope of action is a reference population of nearly 400,000 inhabitants near the capital of the province of Gipuzkoa, Donostia. All the health organizations in Osakidetza are listed in the web application[4].

When creating their account, the volunteer healthcare worker can also indicate whether they are a nurse, a doctor or have any other position, and will mark their specialty or area of care. Even if the healthcare worker indicates his/her specialty, the documents are not displayed considering this specialty, but the one that has been assigned to each pair of documents. We must consider that, for example, some doctors can make extra turns in different specialties. Once the volunteers' details have been indicated, for each pair of reports, the specialty and the document type are indicated. The considered specialties and document types are presented in section 3.1, and listed in the first column of Table X.3 and Table X.2 respectively.

2.3 Web application for corpus collection

4 https://www.osakidetza.euskadi.eus/transparencia-buen-gobierno/-/organigramas-de-organizaciones-de-servicios-de-osakidetza/

For collecting a proper corpus, it is important to provide adequate guidelines to the users. With this aim, the volunteers working on the elaboration of the bilingual corpus were asked to follow these guidelines:

- align the source and target text at sentence level. To do so, write each sentence in one line. If a source sentence needs more than one target sentence to translate, include them in the same line, and vice versa.
- use the most natural way of writing as possible. We need a corpus the most realistic as possible, and so it should be the language used[5].
- use a formal style; lexical variations related to dialectal use are accepted, but not orthographical variations that deviate from standardized terms.

Apart from providing adequate guidelines, one of the aims of the web application is to reduce the effort the volunteers may do writing the records. For that purpose, we included in the web application some tools that may be of help when collecting the bilingual corpus:

- Access to a set of 48 reference discharge records of different specialties in Basque (Joanes Etxeberri Saria V. Edizioa 2014).
- A dynamic bilingual dictionary which the participants can enrich with their collaborations, so anyone can check the proposals of their mates.
- An integrated medical dictionary in the Basque writing textarea, that whenever the volunteer types the colon symbol, it searches the string typed next in a specialized

5 It is well known that the type of language used by healthcare workers is not specially correct. This is mainly due to the fact that they often write quickly, which leads them to make typing mistakes, write sentences that do not follow grammatical rules, and use made up expressions or acronyms.

dictionary. Thus, users avoid checking in a different tab for the source word in a digital dictionary, and speeds the writing process.

In addition, the biggest deal collecting a parallel corpus is the alignment of the source and target sentences. Even if the guidelines in this respect were clear, it should be considered that sometimes a sentence in one language needs more than one sentence in the other language. This may be faced using an automatic sentence splitter, but then some errors could inevitably be introduced. Trying to minimize these, we developed some tools to perform the sentence mapping, by a) splitting the source text when copying to the app, having a sentence by line; and b) underlining with colors the corresponding source and target sentence as seen in Figure X.1.

Figure X.1: Snapshot of the web application for corpus collection, highlighting in colors the aligned sentences in Spanish (left, under '*Gaztelania'*) and Basque (right, under '*Euskara'*), and showing the menus for specialty ('*Espezialitatea'*) and document type ('*Txosten mota'*).

To conclude this section, we would like to highlight two issues concerning the corpus collection: i) the way in which the collection process has been organized and ii) the fact that patient privacy is always guaranteed.

Each OSI has had a project manager who has listed the names of the bilingual healthcare workers who might be interested in participating in Itzulbide. Both the institutional representatives of the health system and the technical managers, and some of the authors of

this chapter, have gone (before the COVID-19 pandemic) to some of these centers to make a presentation of the project, explain the technical part of the corpus collection application and solve professionals' doubts. During the pandemic period, this presentation has been recorded in the form of a video and made available to interested sanitary workers. Each OSI decides whether or not to compensate the healthcare workers for their collaboration, and if this is the case, in what way (with days off, for example).

The medical reports collected in the application do not have to be real, but if they are, they do not contain personal data of the patient (name, place of birth etc.). In Osakidetza, the patient's personal information and the medical reports are stored separately and linked by a code. Once the MT system is implemented in production, it will run on Osakidetza's Graphics Processing Unit (GPU) servers, so the flow of information will be internal.

3 Resources and systems

3.1 Bilingual corpus from Itzulbide

As an intermediate result of the corpora collection part of the Itzulbide project described in section 2, we extracted all the bilingual sentences introduced in the web application until April 21, 2020. Even if the guidelines for aligning the source and target text at sentence level were clear, the number of sentences automatically extracted in each language was not exactly the same, so a manual revision was performed until we had a true parallel corpus aligned at sentence level.

The sentences were originally grouped at document level, and each document had a variable controlled by the user to specify if the document was finished or not. In the most usual scenario, physicians wrote the documents in Spanish as they were working with the patient, and later translated it into Basque with the help of the tools available in the web application designed for the corpus collection[6]. If a document was marked as finished, we considered that it had been properly reviewed by the doctor who wrote it, so we further used this variable to create the evaluation corpus only with sentences coming from documents marked as finished.

Another important variable for classifying the compiled sentences is the document type. We defined 5 different document types as they are commonly distinguished in clinical scenarios: 1) hospitalization reports, written when the patients are initially derived to the hospital; 2) progress reports, indicating the evolution of the patients while they are in the hospital; 3) discharge reports, written at the end of the stay of a patient in the hospital; 4) informative permissions, used when a patient has to go through a surgery or any procedure that involves some risk; and 5) others, for documents not filling any of the above characteristics. Since Osakidetza's main priority is to translate progress reports and discharge reports, we used this variable for selecting the sentences to be used for evaluation purposes in one of our evaluation scenarios defined at the end of this section.

Finally, the documents in Itzulbide are also classified by specialty (e.g.: emergencies, nursing, pediatrics, etc.). This information will be used in some of our defined scenarios for helping

6 We are aware that the fact that most of the doctors and nurses write their health records initially in Spanish, and later translate them into Basque, makes the further evaluation in the Basque-to-Spanish direction suboptimal, but we are constrained by the fact that currently, healthcare workers are forced to write the documents in Spanish; and given the lack of Basque standardized clinical terminology, writing them initially in Basque would require more effort from the physicians while they are dealing with their patients.

both the training process, by using distinctive tags for each specialty, and the evaluation process, measuring the performance of the designed systems in a given specialty.

Table X.1, X.2 and X.3 sum up the statistics of the bilingual corpus from Itzulbide, disaggregated by the above mentioned variables. In all these tables, the number of documents corresponds to the total number of compiled documents, including unfinished documents that could have been written only in one language; while the number of sentences and tokens correspond to the bilingual corpus obtained after automatic extraction and manual alignment. Both in the following Tables and main text, we use “eu” and “es” abbreviations to refer to Basque and Spanish languages respectively, corresponding to the standard ISO 639-2 codes commonly used in natural language processing.

Table X.1: Statistics of Itzulbide bilingual corpus disaggregated by the state of the document

State	Documents	Sentence pairs	Tokens (eu/es)
finished	1,774	23,695	198,503 / 236,462
unfinished	179	2,742	19,569 / 23,034

Table X.2: Statistics of Itzulbide bilingual corpus disaggregated by document type

Type	Documents	Sentence pairs	Tokens (eu/es)
hospitalization report	24	625	4,989 / 5,494
progress report	1,333	15,069	110,699 / 127,036
discharge report	260	4,424	29,660 / 33,174
informative permission	139	3,006	42,193 / 55,639
others	197	3,313	30,531 / 38,153

Table X.3: Statistics of Itzulbide bilingual corpus disaggregated by specialty

Specialty	Documents	Sentence pairs	Tokens (eu/es)
oral and maxillofacial surgery	13	86	782 / 835
oral and maxillary surgery	1	22	323 / 453
anesthesia and resuscitation	22	124	812 / 947
respiratory system	51	2,360	16,893 / 17,620
internal medicine	182	4,392	31,201 / 35,922
digestive system	39	1,183	8,349 / 9,836
short stay psychiatry	7	120	1,381 / 1,507
out-of-hospital emergencies	4	2	9 / 36
nursing	156	1,101	10,537 / 13,413
diagnostic radiology	40	240	1,738 / 2,135
rehabilitation	10	53	321 / 338
ongoing care	51	413	3,827 / 4,276
home hospitalization	76	642	4,077 / 4,685
unknown	104	2,413	29,594 / 38,624
family medicine	251	2,483	17,198 / 19,023
pharmacy	50	795	8,534 / 11,024
gynecology and obstetrics	4	53	315 / 348
cardiology	1	25	213 / 280
general surgery	2	15	128 / 134
health management unit	1	22	467 / 639
emergency department	226	2,940	20,003 / 22,043
intensive care medicine	33	643	4,811 / 5,674
otorhinolaryngology	52	1,019	6,260 / 7,319
pediatrics	74	374	2,804 / 3,267
preventive medicine	1	23	108 / 123
psychiatry	172	746	6,619 / 8,165
trauma	104	1,219	9,054 / 10,950
urology	125	2,751	29,500 / 37,306
palliative care	97	156	1,915 / 2,278
palliative care unit	3	21	286 / 275
management	1	1	13 / 21

3.2 Other bilingual corpora from the health domain

Given the need for big quantities of in-domain data for training state-of-the-art Neural Machine Translation (NMT) systems, our main priority has been to collect as much as possible bilingual corpora from the health domain. To this end, apart from carefully preprocessing the 26,437 bilingual sentences from Itzulbide described in section 3.1, we have compiled 541 more bilingual sentences extracted from 17 clinical cases written in the Basurto hospital (Magnini et al. 2020). These clinical cases are available on the web[7], and specifically, the ones we used were written during 2014 and 2015. Note that these documents are categorized by different specialties than the ones used in the Itzulbide corpus, so these 541 sentences will not be tagged by specialty for the experiment described in the end of this section.

In addition, we have downloaded documents for professional health workers published by Osakidetza[8]. From all the documents available in that website, we omitted the administrative ones (in Spanish: '*Planes y programas anuales y plurianuales*' and '*Memorias Osakidetza*') and only made use of the documents that were available in both Basque and Spanish in the date of download (October 1, 2020). These documents were converted from pdf to text, and later sentence segmentation was performed for the documents in each language, using an in-house program that was also used for the Itzulbide bilingual corpus. Finally, sentences in both languages were manually aligned; and for making the most of the scarce in-domain bilingual

7 https://github.com/hltfbk/E3C-Corpus
8 https://www.osakidetza.euskadi.eus/profesionales/-/publicaciones-profesionales/

corpora, in case one sentence in one language corresponded to many sentences in the other language, these many sentences were joined using ‘;’ as a separator. In this way we obtained 22,051 more parallel sentences from the health domain.

3.3 Clinical terminologies used as sentences

As additional bilingual in-domain data, we have extracted clinical terminologies from different sources. Even if these clinical terms are not actually sentences, considering the rich vocabulary of health domain and the low resources for eu/es language pair, we think that they can be useful for improving the coverage of the NMT system, helping to translate clinical terms that probably do not appear in the few available bilingual sentences. Moreover, since the sentences we want to translate are usually short and often omit verbs, we consider that using clinical terms formed by a few tokens as sentences will not have any negative effect on the final performance of our system.

Most of the clinical terminologies we have used come from the automatic translation into Basque of the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT), which is described as the most comprehensive clinical terminology collection in the world (Perez-de-Viñaspre 2017). The system automatically creates Basque terms corresponding to the original English terms by combining the use of dictionaries, transliteration tools, and rule based systems, and for training our MT models we used all the terms containing up to 11 tokens (896,898 in total). As these terms are automatically created in Basque, there is often

more than one term in Basque for each concept, which we think can be especially useful for translating in the eu-es direction.

Another clinical terminology collection we have used is the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). In our case, we have used the manual translation of the concept descriptions in Basque as provided by the organisers of WMT Biomedical shared task (Bawden et al. 2020), obtaining the corresponding descriptions in Spanish directly from ICD-10. From this process we compiled 27,696 additional segments.

Finally, as an initial step for making the systems ready to translate COVID-19 related terms, we have compiled a few dictionaries from an interim release of SNOMED CT[9], consisting of 84 terms, having the English terms translated into Basque by a translator of Osakidetza. Additionally, we compiled 126 COVID-19 related terms compiled by Elhuyar foundation, including all the terms published until June 18, 2020.

3.4 Bilingual out-of-domain corpora

When looking for out-of-domain bilingual corpora that could be useful for our task of translating clinical notes, we have looked for a balance between compiling as much data as possible (given the lack of resources) and granting a minimum quality (given the desired high accuracy for a sensitive domain like the health domain). In this sense, we have chosen two corpora that have been professionally translated and, whether have been previously tested on eu/es NMT, or they have similar characteristics to the clinical texts we want to translate.

9 https://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release-COVID-19

The first and bigger out-of-domain corpus was originally composed by 4.5M sentences, being half of them a repetition of a corpus from the news domain (Etchegoyhen et al. 2020), and the other half coming from diverse sources such as administrative texts, web-crawling and specialized magazines. This corpus has been previously used for MT of clinical texts between Basque and Spanish (Soto et al. 2019a; Soto et al. 2019b; Soto et al. 2020). For reducing the noise introduced by out-of-domain sentences, we have applied a language identification tool[10] to exclude sentences where most of the terms are named entities like locations or person names. This way, the vocabulary of the out-of-domain corpus is reduced, so a bigger part of the limited vocabulary of the NMT system can be used for translating health domain terms. By removing the sentences that are classified as another language in each of the eu and es sides of the corpus, we filtered 3,703,757 sentences from the original 4.5M sentences.

The other out-of-domain corpus we use is HAC (Sarasola et al. 2015), compiled by OPUS (Tiedemann 2012) and formed by 566,738 sentences coming from the translation of literary books. Even if the domain is very different from the health domain, we chose this corpus for being translated by professional translators and having similar average sentence length to our clinical domain corpus. On the contrary, we discard other publicly available corpora like OpenSubtitles or GNOME for not having the desired translation quality.

Regarding the preprocessing of the out-of-domain corpora, given that the maximum number of tokens in the clinical domain corpus is 98, we removed all sentences longer than 100 tokens using Moses tools[11]. We also tried to remove sentences shorter than 3 tokens or using

10 https://github.com/saffsd/langid.py
11 https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl

punctuation normalization tools, but both experiments decreased the performance of our NMT system as shown in preliminary results. All of the corpora, including the monolingual ones presented in the next section, were tokenized and truecased using Moses tools[12], being the Truecase model learned on the bigger out-of-domain corpus with its original 4.5M sentences.

Table X.4 sums up the statistics of the bilingual corpora described in sections 3.1-4.

Table X.4: Description and statistics of the diverse bilingual corpora

Corpus	Sentence pairs	Tokens (eu/es)
Itzulbide bilingual corpus	26,437	218,072 / 259,496
bilingual sentences from Basurto hospital	541	5,254 / 5,185
Osakidetza's publications for health workers	22,051	299,203 / 350,361
SNOMED CT clinical terms	896,898	3,074,750 / 5,309,227
ICD-10 terms	27,696	229,248 / 175,627
COVID-19 related terms from SNOMED CT	84	579 / 729
COVID-19 related terms from Elhuyar	126	263 / 243
out-of-domain corpus (news and others)	3,703,757	66,284,429 / 95,714,868
HAC corpus (literary)	566,738	8,861,175 / 10,956,345

3.5 Spanish monolingual corpora from the health domain

For the eu-es translation direction we leveraged the Electronic Health Records (EHR) in Spanish already compiled from Osakidetza in previous research projects. For the experiments performed in this work, we have used discharge reports from two hospitals: Galdakao-

12 https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl and https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl respectively

Usansolo and Basurto. The EHRs from Galdakao-Usansolo hospital consist of 142,154 documents compiled from 2008 to 2012, while the discharge reports from Basurto hospital sum up to 57,569 documents written in 2014. After performing sentence segmentation using the same preprocessing applied to the bilingual in-domain corpora, removing repeated sentences in the corpus from each hospital, and deleting sentences only containing the document ID and/or date, we obtained 1,921,672 sentences from Galdakao-Usansolo and 905,893 from Basurto.

Due to privacy issues, these corpora cannot be made publicly available. The documents were given to us without any personally identifiable information, and before using the corpus from each hospital, it was further de-identified by means of shuffling the sentences. Only authors who had previously signed a non-disclosure commitment had access to them.

Table X.5 shows the statistics of the Spanish monolingual corpora from the health domain described in this section. In this table, the number of documents corresponds to the total number of compiled documents; while the number of sentences and tokens correspond to the corpus obtained after filtering and preprocessing.

Table X.5: Statistics of the Spanish monolingual corpora from the health domain

Hospital	Documents	Sentences	Tokens
Galdakao-Usansolo	142,154	1,921,672	32,084,578
Basurto	57,569	905,893	11,812,057

3.6 System training and evaluation

Being NMT the state-of-the-art method for MT, and based on previous work on translation between Basque and Spanish (Etchegoyhen et al. 2018), we use NMT for training our systems. More specifically, taking into account previous work on translation of clinical texts from Basque to Spanish (Soto et al. 2019b), we choose the Transformer (Vaswani et al. 2017) architecture. From the different available implementations, we decided to use Fairseq (Ott et al. 2019), for being the one mostly used in recent MT shared tasks (Bawden et al. 2020). Considering the rich morphology of Basque language and the rich terminology of health domain texts, we use Byte Pair Encoding (BPE) (Sennrich et al. 2015) with 90,000 merge operations for subword segmentation. Additionally, we try BPE-dropout (Provilkov et al. 2020) with 0.1 probability for preprocessing our training corpora, whether applied in both sides of the training corpus or only in the source side. We believe this regularization technique can be especially useful in our rich vocabulary setting, and can improve the robustness of the system against the usual typos appearing in EHRs.

As a general method for domain adaptation, we use regular fine-tuning, using the corpora most similar to our evaluation data for fine-tuning, and the rest for pretraining. In this sense, after extracting the corresponding evaluation corpus from the Itzulbide bilingual corpus (1,000 sentences for validation and 1,000 for testing), we use the remaining 24,437 sentences for fine-tuning. In addition, we also use the 541 bilingual sentences from Basurto hospital described in the first paragraph of section 3.2 for fine-tuning. The rest of the bilingual corpora presented in Table X.4 is used for pretraining.

For both translation directions, we pretrained the systems for 20 epochs when using BPE and for 50 epochs when applying BPE-dropout. In each case, we fine-tuned the systems for the same number of epochs, calculated the BLEU (Papineni et al. 2002) scores on the validation set in each of the models saved after every epoch, and finally calculated the BLEU score on the test set with the model that obtained the highest BLEU on the validation set.

We set two different configurations for evaluating our es-eu models: in the first one, apart from testing the systems preprocessed with different subword segmentation methods, we wanted to evaluate how our system would work in a specialty for which we have no training data. For simulating this scenario, we extracted all the finished sentences from trauma specialty before extracting the evaluation corpus, and used all of these sentences from trauma as a second test set. The sentences corresponding to trauma specialty coming from documents marked as unfinished or from the 541 bilingual sentences from Basurto hospital were not used for training in this scenario. We chose this specialty for having around 1,000 finished sentences in the Itzulbide bilingual corpus and having a distinct terminology to the other specialties. Thus, we will have six different results from this configuration, corresponding to the use of BPE, BPE-dropout and BPE-dropout applied only in the source language for subword segmentation; and evaluated both in the trauma test set and in the test set extracted from the remaining specialties.

In the second scenario, our aim was to develop a system that matched the requirements of Osakidetza as closely as possible, so we used sentences from all the specialties for both training and evaluation, and selected only discharge reports and progress reports for

evaluation, using the sentences coming from other document types only for training. For training this system we used the subword segmentation method that obtained the best results in the previous scenario; and additionally, we tried adding tags for identifying the specialty of the sentences from Itzulbide used for fine-tuning. For doing this, we defined an acronym of 3 characters for each specialty listed in Table X.3, and include it between '<' and '>' marks (for instance, '<DIG>' for digestive system). Then, for each Basque sentence in the Itzulbide bilingual corpus, we inserted the corresponding specialty tag in the beginning of the sentence, separated by a blank space. Specifically, we inserted the tag after applying subword segmentation, so the tags did not influence the learned BPE model. After evaluating the systems with and without using these tags, we used the best performing system resulting from this process for translating the available EHRs in Spanish into Basque. The translation was done using unrestricted sampling (Edunov et al. 2018) as decoding method[13], with a buffer size of 1,000 and a batch size of 50.

The created synthetic corpus was added to the training corpora used in the es-eu direction for training the systems in the eu-es direction. We conducted two experiments using this synthetic corpus: in the first one we added the pseudo-parallel corpus created via back-translation (Sennrich et al. 2016), and in the second one we further added the monolingual corpus as both source and target corpus, using the technique known as copying (Currey et al. 2017). In both experiments, we applied BPE-dropout and pretrained/fine-tuned the systems for 50 epochs.

13 For details about how to implement unrestricted sampling in fairseq, see: https://github.com/pytorch/fairseq/issues/308

4 Results and discussion

4.1 Spanish to Basque translation direction

First, we compare the results in the es-eu translation direction when using BPE, BPE-dropout or BPE-dropout applied in the source side only. Table X.6 presents the BLEU scores obtained with these subword segmentation methods, evaluated in the trauma test set and in the general test set formed by sentences from other specialties.

Table X.6: BLEU scores in the es-eu translation direction using different subword segmentation methods (best results in each test set are marked in bold)

Subword segmentation method	BLEU (general test set)	BLEU (trauma test set)
BPE	36.24	19.28
BPE-dropout	**37.42**	**19.52**
BPE-dropout (source side only)	37.06	18.57

We observe that BPE-dropout obtains the best results in both test sets, so we use this subword segmentation method for further experiments. Regarding the big difference between the BLEU scores obtained in the general test set and the trauma test set, one would think that the performance in the held-out specialty is very low, but after looking to the generated translations we see a different picture: most of the differences between the human translations and the machine translations do not correspond to errors of the automatic translations, but to errors on the manual translations. These differences can be classified in diverse categories,

like typos (‘*egunena*’ instead of ‘*egunean*’, meaning ‘in the day’), orthographic errors (‘*protezi*’ instead of ‘*protesi*’, meaning ‘prosthesis’), or even the use of different dialects of Basque (‘*ondo*’ instead of ‘*ongi*’, meaning ‘good’). In the analyzed 100 sentences we did not observe any error that could be attributed to the domain shift; however, a thorough human evaluation should be performed by healthcare workers to study the effect of using our MT tool to translate sentences from a specialty not seen in the training corpus.

In the second scenario we tested the effect of adding tags to the diverse specialties included in the bilingual corpus from Itzulbide. As stated in the previous section, in this setting we evaluate only on sentences coming from progress reports and discharge reports. Table X.7 shows the BLEU scores obtained with and without specialty tags (for completion, the BLEU score of the system before fine-tuning is 18.28).

Table X.7: BLEU scores in the es-eu translation direction with and without specialty tags (best result is marked in bold)

Use of tags	**BLEU (progress reports and discharge reports)**
without tags	**31.97**
with tags	30.67

We observe that the system without tags performs better, so we use this model for translating the Spanish monolingual corpora from the health domain and do not use tags for the following experiments in the eu-es translation direction.

4.2 Basque to Spanish translation direction

For the eu-es translation direction, we perform two experiments including the Spanish monolingual corpora from the health domain in two different ways: one including the pseudo-parallel corpus resulting from back-translation (Sennrich et al. 2016), and the second one applying also the technique known as copying (Currey et al. 2017). Table X.8 presents the BLEU scores for these settings. For a better analysis, we include the results before and after fine-tuning, and also show the size of the pretraining corpus in each configuration (the fine-tuning corpus is formed by 24,978 sentences).

Table X.8: Number of sentences of the pretraining corpus, along with BLEU scores in the eu-es translation direction after back-translation and further copying; before and after fine-tuning (best result in each case is marked in bold)

Method	**Pretraining sentences**	**BLEU (before fine-tuning)**	**BLEU (after fine-tuning)**
+back-translation	7,956,799	38.50	**50.67**
++copying	10,729,257	**38.74**	49.64

Contrary to previous work that did not use bilingual clinical domain data (Soto et al. 2019a), we observe that in our case the technique of copying does not improve the results, so we present the system using only back-translation as the best performing system.

4.3 Error analysis

In this last subsection, we analyze the best performing systems in both translation directions, which obtained 31.97 BLEU points in es-eu and 50.67 in eu-es. Before analyzing the generated translations, note that BLEU metric underestimates the translation performance into morphologically rich languages as Basque; so, even if it is expectable that the eu-es system performs better thanks to the use of Spanish monolingual EHRs, part of the big gap between the BLEU scores in es-eu and eu-es can also be attributed to this fact.

As a first simple analysis of the generated translations, we checked if numbers were correctly translated, given the special relevance it could have to incorrectly translate a result of a medical test, a prescribed dose, an appointment date, hour, etc. With this aim, we manually checked how numbers were translated in the first 100 sentences of the test set in both translation directions, and confirmed that all numbers were translated correctly.

Secondly, given the lack of standardized clinical terminology in Basque, we wanted to analyze how clinical terms were translated in the es-eu direction. For doing this semi-automatically, we used the 896,898 bilingual terms coming from the automatic translation of SNOMED CT into Basque as a reference, looking for their appearance in both sides of our test set translated by the best performing system. Since the SNOMED CT terms in Basque were automatically created, this analysis is conditioned by the way this term creation was done. That is why, when looking for these terms in the test set, we observed that all the clinical terms identified in both Basque and Spanish corresponded to terms that were whether identically written or transliterations, as this was one of the methods used for automatically creating the Basque terms. Specifically, we detected 44 bilingual terms in 37 sentences out of

the 1,000 sentences of the test set, and observed that all of them were translated correctly. The only differences between the terms included in the generated translations and the automatically created SNOMED CT terms in Basque corresponded to one appearance of the acronym '*EKG*' instead of the extended '*elektrokardiograma*' (meaning 'electrocardiogram') and three appearances of '*arnas-hestu*' or the shrinked '*arnasestu*' instead of the recommended '*disnea*' (meaning 'dyspnea'). In the latter case, we can consider that the generated translations ('*arnas-hestu*' / '*arnasestu*') are more informal than the more technical '*disnea*', but we can anyways accept them as correct translations for being similar to the 'breathless' term found in SNOMED CT as synonym of 'dyspnea' ('*arnas*' meaning 'breath' and '*hestu*' meaning 'tight').

5. Conclusions and future work

Analyzing the automatic evaluation results shown in the previous section, we can conclude that the developed systems obtain good results in both translation directions, especially in the eu-es direction thanks to the big number of EHRs available in Spanish (around 2.8M sentences). We also observe that using a bilingual in-domain corpus for fine-tuning, even if limited to 24,978 sentences, greatly improves the final performance of the systems, boosting 14 BLEU points when translating from Spanish into Basque, and 12 from Basque into Spanish. This validates our effort of compiling the bilingual clinical domain corpus as a necessary first step for developing a domain adapted MT model.

Regarding the translation of clinical terminologies, we have observed that our systems always generate correct terms, and most of the time use the terms coming from the previous automatic translation into Basque of SNOMED CT. However, we believe that it would be helpful for future work to have a standardized clinical terminology in Basque, so we could properly evaluate the MT of clinical terms.

Regarding data preprocessing, we have proved that the use of a regularization technique like BPE-dropout improves the results in a domain where typos, misspellings, etc. are usual, giving us an extra BLEU point in the es-eu translation direction. Additionally, we have tried using tags for identifying the specialties in the clinical domain bilingual corpus, but saw no gains from including them as a way to guide the NMT system, probably due to the limited number of training examples for some of the specialties.

Overall, even if the automatic evaluation scores and error analysis show promising results, a human evaluation should be performed before implementing these systems for translating clinical texts in a real-life scenario. It would also be interesting to extend this human evaluation for testing the performance of the models in a held-out specialty, as we did in this work by automatic means. We leave this human evaluation as future work.

Furthermore, we plan to keep compiling bilingual/monolingual clinical domain corpora to continuously improve the performance of our systems, focusing on the eu-es translation direction. With this aim, we plan to try diverse strategies for back-translation (Caswell et al. 2019; Graça et al. 2019; Hu et al. 2019) and apply data selection methods to the back-translated data (Soto et al. 2020).

Up to now, we have developed a useful MT tool, obtaining 50.67 BLEU points, that could help Spanish monolingual speakers understand the clinical notes written in Basque. However, in case this system is implemented in a real-world scenario, given the critical effects it could have a decision based on an incorrect translation, we recommend that a bilingual speaker reviews the generated translations and corrects them if necessary. In any case, the deployment of this model can encourage Basque speaking healthcare workers to write their reports in Basque, which is an objective of the Itzulbide project that has already started to be fulfilled with the 26,437 sentences used for building this initial MT tool. Furthermore, these clinical domain corpora can also be helpful for other NLP tasks.

References

Bawden, R., G. Di Nunzio, C. Grozea et al. 2020. Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional Languages. In *5th Conference on Machine Translation (WMT2020)*.

Caswell, I., C. Chelba, and D. Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (WMT2019)*.

Currey, A., A. V. Miceli-Barone, and K. Heafield. 2017. Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation (WMT2017)*.

Desjardins, L. 2003. *La santé des francophones du Nouveau-Brunswick*. Petit-Rocher, Société des Acadiens et des Acadiennes du Nouveau-Brunswick.

Edunov, S., M. Ott, M. Auli, and D. Grangier. 2018. Understanding back-translation at scale. arXiv preprint arXiv:1808.09381

Etchegoyhen, T., E. Martínez, A. Azpeitia et al. 2018. Neural Machine Translation of Basque. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*.

Etchegoyhen, T., and H. Gete. 2020. Handle with Care: A Case Study in Comparable Corpora Exploitation for Neural Machine Translation. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*.

European Observatory on Health Care Systems. 1999. Luxembourg: Health system review. *Health Systems in Transition*.

Gerkens, S., and S. Merkur. 2010. Belgium: Health system review. *Health systems in transition* 12(5):1-266.

Graça, M., Y. Kim, J. Schamper, S. Khadivi, and H. Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (WMT2019)*.

Hu, J., M. Xia, G. Neubig, and J. Carbonell. 2019. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of The 57th Annual Conference of the Association for Computational Linguistics (ACL 2019)*.

Joanes Etxeberri Saria V. Edizioa. 2014. *Donostia unibertsitate ospitaleko alta-txostenak*. Komunikazio Unitatea, Donostiako Unibertsitate Ospitalea.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2020. The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. In Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020).

Ott, M., S. Edunov, A. Baevski et al. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Papineni, K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*.

Perez-de-Viñaspre, O. 2017. Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque. Ph.D. thesis, University of the Basque Country.

Provilkov, I., D. Emelianenko, and E. Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*.

Sarasola, I., P. Salaburu, and J. Landa. 2015. *Hizkuntzen Arteko Corpusa (HAC)*. University of the Basque Country UPV/EHU (Euskara Institutua).

Sennrich, R., B. Haddow, and A. Birch. 2015. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909

Sennrich, R., B. Haddow, and A. Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Soto, X., O. Perez-de-Viñaspre, G. Labaka, and M. Oronoz, (2019a). Neural machine translation of clinical texts between long distance languages. *Journal of the American Medical Informatics Association* 26(12):1478-1487.

Soto, X., O. Perez-de-Viñaspre, M. Oronoz, and G. Labaka. (2019b). Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*.

Soto, X., D. Shterionov, A. Poncelas, and A. Way. 2020. Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of The 58th Annual Conference of the Association for Computational Linguistics (ACL 2020)*.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*.

Vaswani, A., N. Shazeer, N. Parmar et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*.

Comparing and combining tagging with different decoding algorithms for back-translation in Neural Machine Translation: an analysis from a lexical diversity perspective

Xabier Soto, Olatz Perez-de-Vinaspre, Gorka Labaka, Maite Oronoz

Manuel Lardizabal 1, Donostia

Abstract

Back-translation is a well established approach to improve the performance of Neural Machine Translation (NMT) systems when large monolingual corpora of the target language and domain are available. Recently, diverse approaches have been proposed to get better automatic evaluation results of NMT models using back-translation, including the use of sampling as decoding algorithm for creating the synthetic corpus. Alternatively, it has been proposed to append a tag to the back-translated corpus for helping the NMT system to distinguish the synthetic bilingual corpus from the authentic one. However, it is still not clear which is the best approach for developing a given NMT system, and most importantly, not all the combinations of the previous approaches have been tested. In this work, we empirically compare and combine existing techniques for back-translation in a realistic low resource setting: the translation of clinical notes from Basque into Spanish. Apart from automatically evaluating the MT systems, we analyze the different synthetic corpora by measuring their lexical diversity (LD), and study the gender bias of typically gender-stereotyped words. For reproducibility and generalizability, we repeat our MT and LD experiments for German to English translation using bilingual data from WMT Biomedical shared task and discharge reports in English extracted from MIMIC III. The results suggest that in lower resource scenarios tagging only helps when using sampling for decoding, in contradiction with the previous literature using cor-

pora from the news domain. When fine-tuning with a few thousand bilingual in-domain sentences, tagging also helps in a restricted sampling scenario, improving the MT scores of tagged beam search or unrestricted sampling. We will provide all the scripts for preprocessing, training and evaluation as supplementary material, which could be useful for advancing the research on lower resource language pairs and domains.

Keywords: Neural Machine Translation, back-translation, decoding algorithms

1. Introduction

Neural Machine Translation (NMT) [1, 2, 3] is the state-of-the-art approach for developing Machine Translation (MT) systems. However, as NMT is based on artificial neural networks, its performance is dependent on big quantities of bilingual sentences, which are not available for all language pairs and domains.

Back-translation (BT) [4], based on the automatic translation of a corpus from the target language into the source language for augmenting the training data, has become a de facto standard for improving the performance of NMT models, provided that large monolingual corpora in the target language and domain are available.

When producing an output sentence, MT systems have to implement an efficient technique that avoids looking for all the possible output sentences and choosing the one with the highest probability according to the distribution of the training data. Typically, beam search [5] is used for generating both the output sentences of NMT systems and the synthetic sentences produced by BT systems.

Recently, [6] proposed to use sampling for BT as one way to further improve the performance of NMT systems. Specifically, their approach of randomly sampling from the output distribution obtained the best results on average comparing to other decoding algorithms, including beam search.

On the contrary, [7] suggest that the improvement derived from using sampling for BT comes from the fact that the final NMT system can identify the

synthetic corpus for having been generated by sampling instead of beam search, so they propose a simple alternative consisting of adding a tag to the corpus generated by the BT system using traditional beam search. They also tried to tag the output of the BT system using noising as proposed by [6], but not the one using sampling.

Concurrent work by [8] instead propose some variations to the sampling approach, consisting of disabling the label smoothing option when training the BT system, and restricting the sampling by setting a minimum value to the probability of the output sentences or limiting it to the top-k values. From these options, the last one obtained the best results, which we refer to as restricted sampling.

Thus, we would have six options for generating the BT corpus, depending on which decoding algorithm is used, and whether tagging is used or not:

1. beam search (before [6])
2. unrestricted sampling [6]
3. restricted sampling [8]
4. tagged beam search [7]
5. tagged unrestricted sampling (this work)
6. tagged restricted sampling (this work)

We compare these 6 methods both in terms of automatic evaluation of NMT systems, and lexical diversity of the synthetic corpora created by the BT systems. For MT automatic evaluation we use BLEU [9], TER [10], chrF [11], and METEOR [12]; while for lexical diversity we measure MTLD [13], TTR [14] and Yule's I [15].

TTR, standing for Type-Token Ratio, is the most common measure for lexical diversity. Its value is obtained by dividing the number of types —defined as the number of different words— by the total number of tokens or words in a given corpus. While easy to interpret, TTR is limited in the sense that their values differ significantly when changing the corpora size, thus it is only a valid metric for comparing lexical diversity of similar sized corpora.

Yule's I is the reversion of Yule's K, or "characteristic constant", which represents the variability of the lexical frequency as the analysed text from the corpus under study gets bigger. Yule's I and Yule's K are thought to be less sensitive to changes in the corpora size. However, both TTR and Yule's I are considered as better suited for small sized corpora.

MTLD or Measure of Textual, Lexical Diversity, sequentially measures the mean length of subsequent n-grams that have the same TTR value. As it is measured sequentially, it is less prone to changes in the values measured on different sized corpora, and it is considered as the most representative metric for measuring the lexical diversity of big corpora as the ones used in MT.

For complementing the LD analysis in the Basque-to-Spanish scenario, we provide the number of appearances of the clinical terms 'patient', 'doctor' and 'nurse', which are marked by gender in the target language but not in the source language, allowing us to measure the gender bias of our corpora/systems.

Finally, we report an estimation of the carbon footprint produced when developing our systems, which can be considered for deciding which approach to take in future works.

2. Related Work

Apart from the works mentioned in the introduction proposing different methods for decoding or tagging the synthetic BT corpus [6, 8, 7], there is some other previous work on comparing different systems for BT.

Probably the most relevant work in this respect is the one that compares different techniques (i.e.: rule-based, statistical or neural MT) for generating the synthetic BT corpus. In this area, the work by [16] firstly compared the use of statistical (SMT) and neural (NMT) systems for BT, without observing significant differences. More similarly to our work, [17] tried rule-based (RBMT), SMT and NMT for BT applied to the translation of clinical texts, obtaining better results with NMT, and specifically the Transformer architecture [18].

[19] went one step further and not only compared the performance of different

techniques for BT, but combined the synthetic corpora created by SMT and NMT systems, probing that the combination of the outputs of both systems was useful. Furthermore, [20] compared and combined the outputs of RBMT, SMT and NMT systems for BT, also analysing the lexical diversity of the generated corpora. They observed that the combination of all systems was in general better than using the output of only one system, and tried to improve the performance by applying data selection [21, 22] to the BT corpus, conditioned on the measured MT and LD metrics for each of the BT systems.

Regarding the use of tags for identifying the BT corpus, [23] concluded that it was advisable to add a tag when the origin of the text was unknown, since systems using BT without a tag overfitted to the synthetic corpus, and even shown to be detrimental when used to translate text originally written in the source language.

Finally, our analysis of the lexical diversity of the BT data generated by different MT systems follows the work of [24], where the authors study the loss of lexical diversity of a given corpus after being translated with SMT and NMT systems. Therefore, in our work we measure the lexical diversity of the BT corpora according to the same metrics they calculate. The linguistic analysis of the gender bias associated to our corpora/systems is also inspired by the work of [25].

3. Material and methods

We test the six methods presented in the introduction for a real use case: the translation of clinical notes from Basque to Spanish (eu-es). This work is part of an ongoing project that aims to implement an MT system in the Basque public health service (Osakidetza), so Basque speaking healthcare workers can write their reports in Basque without compromising the safety of their patients.[1]

[1] It is expected that the output of the MT system will be post-edited by bilingual experts from Osakidetza before making it available to patients or Spanish monolingual healthcare workers.

The first step in this project is the compilation of a Basque/Spanish (eu/es) parallel corpus of health records to be used for fine-tuning and evaluation, while previously collected Spanish monolingual corpora will be used for BT. Since these corpora are private (see Section 5 for details about privacy), we reproduce our experiments in a similar setting for translating biomedical texts from German to English (de-en), using only publicly available data.

For both language pairs, we preprocess our corpora by tokenizing and truecasing through Moses tools.[2] Further, we apply BPE [26] for 90,000 (eu/es) and 40,000 (de/en) iterations.

For training all our systems, we use the Transformer architecture as implemented in Fairseq [27]. Specifically, we try the hyperparameters proposed in the end of this blog post,[3] training the systems for 30 epochs, and using the 'fp16' option to fasten the training process.

As an exception, for the es-eu system we apply BPE-dropout [28] and train the systems for 50 epochs, as this setting obtained better results on preliminary experiments. In the future, we plan to do the same for the best performing eu-es systems. For de/en, we opt to use regular BPE for better reproducibility.

In the following subsections, we describe the data used for each language pair.

3.1. eu-es corpora

In the eu-es scenario we define four types of data: 1) out-of-domain bilingual sentences, 2) bilingual clinical terms, 3) bilingual clinical notes, and 4) monolingual clinical texts in Spanish. We use the sets 1-3 to train the BT system (es-eu), and later train the final eu-es systems adding the monolingual corpora through BT.

[2] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl` and `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl` respectively

[3] `http://cslab.org/blog/fairseq-basics`

In both translation directions, we apply regular fine-tuning, dividing the training process in two steps: 1) pretraining, using all except the bilingual clinical notes, and 2) fine-tuning, continuing the training of the pretrained systems with the bilingual in-domain sentences. In this case, we pretrain+fine-tune the systems for 30+30 epochs.

Table 1 sums up the domain, languages, number of sentences and use of each of our corpora.

Domain	**Languages**	**Sentences**	**Use**
out-of-domain	eu/es	4,896,719	pretrain
clinical terms	eu/es	924,804	pretrain
clinical notes	eu/es	28,602	fine-tune
clinical texts	es	4,946,293	back-tr.

Table 1: Characteristics and use of the eu/es corpora.

In the following lines, we present some of the details of the training corpora, as enumerated in the beginning of this subsection.

3.1.1. out-of-domain bilingual sentences

In this work we use around 5M sentences of diverse domains, being around 3M sentences the concatenation of a 3 times repetition of a news corpus from the Basque public broadcast service EiTB [29] and a more recent one from the same source [30]. The remaining 2M sentences are from different domains as administrative (IVAP), consumer magazines (Eroski), online magazines (Irrika), translation memories (EIZIE), movie synopses, web crawling [31] and literature [32].

We include as out-of-domain data the sentences extracted from documents published in Osakidetza's website, since their domain is not close to the clinical notes focus of our study. These documents are available online,[4] and were downloaded on October 1, 2020, omitting the administrative ones (in Spanish:

[4] `https://www.osakidetza.euskadi.eus/profesionales/-/publicaciones-profesionales/`

'*Planes y programas anuales y plurianuales*' and '*Memorias Osakidetza*').

3.1.2. bilingual clinical terms

For adapting the pretraining systems to the clinical domain, we leverage clinical terminology available in Basque and Spanish. Most of the 0.9M bilingual terms come from the automatic translation of SNOMED CT into Basque [33], while another 30,000 are manual translations into Basque of ICD-10 concept descriptions in Spanish made available for the WMT Biomedical shared task [34].

Finally, around 200 terms related to the COVID-19 pandemic are compiled, coming around half of them from an interim release of SNOMED CT,[5] and being translated into Basque by a translator of Osakidetza. The remaining terms were collected by Elhuyar foundation and published in their website.[6]

3.1.3. bilingual clinical notes

For fine-tuning and evaluation, we use the bilingual corpus compiled in the project with Osakidetza, where Basque speaking healthcare workers volunteered writing their clinical notes in Basque and Spanish.

These sentences are classified among 5 types: 1) discharge reports, 2) progress reports, 3) hospitalization reports, 4) informative permissions and 5) others. Since the main aim of Osakidetza is to translate discharge and progress reports, only sentences coming from these document types are used for evaluation.

The documents were written by professionals of different specialties (e.g.: pediatrics), from where 2,000 sentences were reserved half for validation and another half for testing purposes. The remaining 28,602 were used for fine-tuning.

[5] `http://www.snomed.org/news-and-events/articles/march-2020-interim-snomedct-release%2DCOVID-19`

[6] This page is currently unavailable, but we can make the term list available upon permission from Elhuyar.

3.1.4. monolingual clinical texts in Spanish

In addition to the collected bilingual data, from previous projects developed with Osakidetza we had access to discharge reports from Galdakao-Usansolo hospital, adding up to around 2M non-repeated sentences; as well as discharge (1M) and progress (2M) reports from Basurto hospital.

Both the bilingual and monolingual corpora from Osakidetza were provided to us without any personally identifiable information (names, surnames, etc.), and it was further de-identified by shuffling the sentences coming from each source. The authors had to sign a non-disclosure commitment before getting access to this private data.

3.2. de-en corpora

For generalization and reproducibility purposes, we also perform our experiments using available data in de-en, as well as clinical notes in English for BT. The bilingual data is the same used for training the baseline systems in the WMT Biomedical shared task [34], consisting of around 3M sentences extracted from the UFAL corpus[7] after removing the "Subtitles" subset. For evaluation we use Khresmoi,[8] also used in [34], where 500 sentences are defined for validation and 1,000 sentences for testing.

For evaluation, and when generating the synthetic corpus through beam search, we use a beam size of 16.[9] This value, along with the 40,000 BPE iterations mentioned above, were optimized for the en-de language pair in [34].

Finally, for BT we use the discharge reports in English available in Mimic III [35].[10] After removing the headers containing unnecessary information, deleting the tags for identifying dates, and erasing the empty lines, this monolingual corpus is reduced to around 2M sentences. We choose to not perform sentence splitting to avoid introducing errors associated with this process. As a conse-

[7] `https://ufal.mff.cuni.cz/ufal_medical_corpus`

[8] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122`

[9] Beam size is 10 for evaluation in the eu/es language pair.

[10] `https://mimic.physionet.org/gettingstarted/access/`

quence, before translating this corpus we filter out the sentences longer than 1,000 BPE (sub)words using Moses cleaning corpus tool.[11] Note that, although there are longer sentences in the training corpus, fairseq skips by default all the sentences longer than 1,024 tokens, so the maximum sentence length of the training corpus is similar to the one of the monolingual corpus used for BT. All the necessary scripts for preprocessing the UFAL, Khresmoi and Mimic III corpora will be included as supplementary material.

3.3. Guidelines for reproducing the de-en experiments

Considering that the eu/es experiments cannot be reproduced for colliding with privacy issues, here we provide some guidelines for the training of the BT (en-de) and final NMT (de-en) systems, including the creation of the synthetic corpora by different decoding algorithms, using or not a tag for identifying the BT sentences.

The first step for reproducing our de/en results would be to download the UFAL, Khresmoi and Mimic corpora, and preprocess each of them by running the corresponding scripts from the supplementary material. Note that all of the provided scripts start with a variable definition section in which the user must specify the directories for the downloaded corpora, as well as the external preprocessing tools for tokenizing, truecasing, corpus cleaning and word segmentation.

After downloading and preprocessing the UFAL and Khresmoi corpora, the en-de system can be trained by running the corresponding scripts calling to the fairseq commands for preprocessing and training.

Once the en-de system is trained, the best checkpoint is selected by measuring the BLEU score for each epoch on the validation set running the fairseq-generate script, and then the model that obtains the highest BLEU score on the validation set is evaluated through MT and LD metrics measured on the test

[11] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/training/clean-corpus-n.perl`

set.[12]

Later, the monolingual (en) Mimic corpus is translated by the en-de system using different decoding algorithms (beam search, unrestricted sampling and restricted sampling), and adding a '<BT> ' tag in the beginning of each sentence back-translated by each of the decoding algorithms.

Finally, the 6 de-en systems are trained by concatenating the bilingual corpora to the synthetic corpora created by each of the different decoding algorithms, whether using a tag or not.

All of the necessary scripts for preprocessing and training will be included in the supplementary material.

4. Results and discussion

4.1. MT automatic evaluation

Table 2 presents the MT automatic evaluation scores of the es-eu and en-de systems used for back-translating the monolingual corpora from the clinical domain. Note that both target languages Basque and German are morphologically richer than the corresponding source languages, so metrics like BLEU, based on word-level accuracy, underestimate the actual MT quality comparing to the same systems trained in the opposite direction ('pretraining+fine-tuning' for eu-es and 'pretraining' for de-en in Table 3).

	BLEU↑	TER↓	METEOR↑	CHRF↑
es-eu	33.88	49.27	47.02	61.02
en-de	29.96	52.63	47.64	60.60

Table 2: MT scores of the back-translation systems.

Table 3 shows the MT evaluation scores of the final eu-es and de-en systems. The first rows for each language pair present the results before adding the BT

[12]The scripts for MT and LD evaluation were developed by other researchers, so cannot be shared but will be made available upon contact with the first author and permission from the authors.

corpus, while the next lines present the values obtained when applying each of the decoding algorithms tested in this work, whether using tagging or not. In the case of eu-es, we include the scores before and after fine-tuning.

	System	BLEU↑	TER↓	MET.↑	CHRF↑
EU-ES	pretraining	26.99	58.61	47.70	53.35
	+fine-tuning	46.67	38.74	63.56	66.46
	+BT (beam search)	44.11	41.54	61.48	66.24
	+fine-tuning	51.37	35.15	67.11	70.10
	+BT (tag. beam search)	41.29	44.45	59.47	64.22
	+fine-tuning	51.99	34.96	67.27	70.11
	+BT (unr. sampling)	43.48	41.39	61.36	65.94
	+fine-tuning	52.68	33.84	67.93	71.06
	+BT (tag. unr. sampl.)	42.07	44.33	59.97	65.13
	+fine-tuning	52.42	34.75	67.51	70.72
	+BT (res. sampling)	44.69	40.83	62.23	66.85
	+fine-tuning	52.90	33.96	68.23	71.12
	+BT (tag. res. sampl.)	42.13	43.71	60.22	65.40
	+fine-tuning	**53.10**	**33.55**	**68.30**	**71.34**
DE-EN	pretraining	42.34	38.55	39.91	67.93
	+BT (beam search)	**44.67**	**37.46**	**40.97**	**69.62**
	+BT (tag. beam search)	44.40	37.63	40.79	69.41
	+BT (unr. sampling)	42.47	41.17	39.58	67.65
	+BT (tag. unr. sampl.)	43.14	38.42	40.35	68.59
	+BT (res. sampling)	40.03	45.73	38.60	66.42
	+BT (tag. res. sampl.)	43.27	38.28	40.51	68.68

Table 3: MT scores of the final eu-es and de-en systems

Beyond the scope of this work, we want to start highlighting that for the eu-es direction, fine-tuning with less than 30,000 sentences (row 2) obtains higher improvements than any of the BT methods (rows starting with '+BT') tried in this work, with the only exception of the chrF value for restricted sampling.

Focusing on the methods under study after applying fine-tuning, we observe

that one of the new combinations tried in this work, tagged restricted sampling, obtains the best scores according to all the MT metrics in the eu-es direction, closely followed by restricted sampling and unrestricted sampling, inverting the order of these two according to TER.

Looking to the generated translations, we see that, regardless of the decoding algorithm, the systems before fine-tuning and not using tagging hallucinate '</- ... -/>' style marks when translating sentences corresponding to typical headers like 'CURRENT DISEASE' or 'TREATMENT'. Analyzing the training corpora, we detect this kind of marked headers in the reports coming from Basurto Hospital, so we will remove these tags in future developments. However, we want to highlight that, not only fine-tuning with clean bilingual data, but also tagging the BT corpora, had the effect of removing this particular noise.

Regarding the de-en direction, where, conditioned by the privacy of clinical data, the size of the training corpora is smaller than for the eu-es counterpart, traditional beam search still obtains the best results, followed by tagged beam search. Most interestingly, we see that, in this particular setting, the effect of tagging is only beneficial when using sampling for BT, complementing the hypothesis of [7], that presents tagged back-translation as a "simpler alternative to noising". With these results, we show that both tagging and sampling can be complementary in lower resource settings.

For complementing the de/en MT scores calculated in biomedical data from Khresmoi, we test these same systems with clinical data from HimL,[13] to analyze possible distortions by the slight domain mismatch between the bilingual biomedical data from WMT Biomedical shared task and the monolingual clinical data from MIMIC III. For converting the HimL data from .sgm to raw text we use the tool available on Nematus.[14] Later we tokenize, truecase and apply BPE as done for the rest of the de/en data. Table 4 presents the results on

[13] http://www.himl.eu/test-sets

[14] https://github.com/EdinburghNLP/nematus/blob/master/data/strip_sgml.py

HimL.[15]

	System	BLEU↑	TER↓	MET.↑	CHRF↑
	en-de pretraining	24.71	59.50	41.06	52.30
DE-EN	pretraining	32.39	50.96	33.52	55.95
	+BT (beam search)	**33.58**	**49.93**	**34.96**	**57.89**
	+BT (tag. beam search)	33.31	50.01	34.36	57.29
	+BT (unr. sampling)	28.70	59.68	31.36	53.12
	+BT (tag. unr. sampl.)	32.42	51.23	33.89	56.42
	+BT (res. sampling)	29.04	58.71	31.90	54.12
	+BT (tag. res. sampl.)	33.31	50.26	34.40	57.06

Table 4: MT scores of the de/en systems on HimL

We observe that beam search also obtains the best results on HimL data in the de-en direction, again followed by tagged beam search for BLEU, TER and chrF, being the results of tagged restricted sampling equal to the latter according to BLEU, and slightly better in terms of METEOR. The main difference comes from the worst results obtained by unrestricted sampling, which in this setting achieves the lowest scores according to all metrics, confirming our hypothesis that unrestricted sampling only works with big corpora.

4.2. LD derived from BT

Table 5 presents the LD values measured on the BT corpora created by each of the methods under study, including the results on the original monolingual corpora for reference. Notice that Yule's I and TTR values are multiplied by 100 for an easier reading.

Comparing the results on each language, we surprisingly see that the MTLD values increase when adding a tag to the BT corpus, while Yule's I and TTR metrics follow our intuition and decrease when adding the same prefix to each

[15]Specifically, on the 1044 sentences coming from the NHS subset, since the remaining sentences from Cochrane are used for validation purposes.

Language	Corpus	MTLD	Yule's I	TTR
es	original	13.99	0.668	0.438
eu	BT (beam search)	13.71	0.863	0.578
	BT (tag. beam search)	14.72	0.799	0.387
	BT (unr. sam.)	13.99	7.628	65.22
	BT (tag. unr. sam.)	14.84	7.123	41.69
	BT (res. sam.)	13.73	2.545	5.851
	BT (tag. res. sam.)	14.72	2.359	3.748
en	original	14.14	0.347	0.129
de	BT (beam search)	14.50	0.899	0.754
	BT (tag. beam search)	15.37	0.841	0.521
	BT (unr. sam.)	15.15	8.376	93.62
	BT (tag. unr. sam.)	15.86	7.890	62.19
	BT (res. sam.)	14.39	3.374	12.64
	BT (tag. res. sam.)	15.15	3.167	8.566

Table 5: Lexical diversity scores of the monolingual corpora before and after BT using different decoding algorithms, whether tagging or not. Yule's I and TTR values are multiplied by 100 for improved readability.

sentence coming from BT. Focusing on the more linguistically relevant LD scores without tagging, we observe that, as expected, unrestricted sampling obtains the highest scores in each language for all metrics. By definition, translations generated through restricted sampling are less diverse than the ones produced by unrestricted sampling, so a human MT evaluation is needed in the eu-es direction to see if the higher MT scores for restricted sampling correspond to an actual increase on MT quality or, as it happens with beam search, these higher MT scores are an artifact of automatic metrics that use to overestimate systems that tend to output more frequent words.

4.3. Preliminary human evaluation

Before carrying out a proper human evaluation by the same healthcare workers who compiled the bilingual clinical eu-es data, we make a first estimation by asking a bilingual biomedical expert to blindly evaluate the quality of the 3 systems that obtained higher MT automatic scores in the eu-es setting, namely 1) tagged restricted sampling, 2) restricted sampling and 3) unrestricted sampling.

For assessing the quality of these systems we focus on the adequacy of the generated translations, comparing their semantics with the ones of the corresponding source sentences and checking the reference translations in case of doubt. Table 6 shows the number of sentences from the first 100 non-repeated sentences of the test set identified as totally correct in terms of meaning for each of the best performing systems in the eu-es direction.

tag. res. sam.	res. sam.	unr. sam.
83	75	83

Table 6: Number of sentences perfectly translated from the first 100 non-repeated sentences of the test set for each of the best ranked systems in the eu-es direction.

We clearly observe that restricted sampling, which obtained the second best MT automatic scores but the lowest LD scores, gets significantly lower adequacy scores (75/100) in this preliminary human evaluation, while tagged restricted sampling and unrestricted sampling obtain the same number of totally correct translations (83/100). This confirms our intuition that, in the absence of a human evaluation, LD metrics can be used to assess the MT quality of different systems trained with the same corpus. Considering this, in the future we will complete this human evaluation by assessing in detail the performance of the best systems according to this preliminary assessment, i.e.: tagged restricted sampling and unrestricted sampling.

4.4. Gender bias

As an example of LD loss, we study the possible gender bias of our corpora/systems. Considering that gender is not marked in Basque but it is marked

in Spanish, we analyze the number of appearances of the nouns associated with typical roles in a clinical scenario ('patient', 'doctor' and 'nurse'), classifying them according to the gender associated with the article accompanying the noun ('*el*' for the masculine form, and '*la*' for the feminine form) and the corresponding suffix '*-o*' or '*-a*' for the Spanish terms '*médico*'[16] and '*enfermero/a*'. By analyzing the outputs of the eu-es systems when each of the above terms appear, we can make a first estimation of the gender bias of our corpora/systems. Table 7 shows the number of appearances of each of the terms described above, disaggregated by the appearances in the train, validation and test sets.

Term (below) / Set (right)	train	validation	test
el paciente	129	1	4
la paciente	120	6	5
el médico	65	0	0
la médico	5	0	0
el enfermero	1	0	0
la enfermera	10	0	0

Table 7: Number of appearances in the train, validation and test sets of each of the terms 'patient', 'doctor' and 'nurse', in their most typical Spanish forms for masculine and feminine genders.

We observe that the term 'patient' ('*paciente*') appears in our corpora in a similar number of sentences regardless of the associated gender, so it serves us as a control variable to show that there is not a preferred masculine/feminine form when the noun is not stereotyped. On the contrary, for the highly stereotyped 'doctor' ('*médico*') and 'nurse' ('*enfermero/a*'), both appear significantly more times in their most stereotyped forms. Specially striking is the case of the term 'doctor', considering that most of the doctors in Osakidetza are women. We

[16]Even if the feminine form accepted by the Spanish language academy is '*médica*', we only find it once in the training set and once in the validation set, so could not use it for analysing the translations of the test set.

will study the possible origin of this bias, but our hypothesis is that one of the sources could be the informative permissions used for training the systems, where the term 'doctor' is used without referring to a specific person.

Analyzing the generated MT outputs, the term 'patient' appears between 3 to 6 times for each of the systems and genders under study, with a maximum difference of 2 appearances between genders for each system; so even if the sample is small, we can say that the system is not gender-biased when translating the word 'patient'. Even if the terms 'doctor' and 'nurse' did not appear in the test set, the term 'nursery' ('*erizaintza*'), appearing in one sentence, was translated as the stereotipally gendered '*enfermera*' by 4 out of 7 systems, the rest keeping the original gender-neutral term 'nursery' ('*enfermería*'). Interestingly, these 3 systems correspond to 3 of the 4 best performing MT systems, excluding the one using restricted sampling, which obtained one of the lowest LD scores. However, a more thorough study should be performed to study the gender bias of the systems using different methods for BT.

4.5. Carbon footprint

To conclude this section, answering to the call made by [36], we report the carbon footprint derived from training our systems. For doing that, we obtain the training times from the log files for each system, accordingly calculate the consumed power, and then estimate the corresponding CO_2 emissions.

Table 8 shows the measured time, power consumption and CO_2 emissions estimated for each of the developed systems. Each experiment was done using a single Nvidia Titan V GPU with a maximum power of 250W. We estimate the CO_2 emissions by applying equations (1) and (2) in [36], considering only the power consumed by our GPUs. Note that the training of the es-eu system is done for 50 epochs, while the rest are performed for 30 epochs.

For interpreting these results, it must be considered that the default implementation of fairseq is not optimized to use the maximum power of the GPUs at any time, so the presented values must be taken with caution as a clear overestimation. We leave as future work modifying the fairseq hyperparameters to make

System	Time (h)	Power (kWh)	CO_2e (lbs)
es-eu	81.93	32.36	30.88
eu-es	38.66	15.27	14.57
eu-es + BT (b.s.)	71.90	28.40	27.10
eu-es + BT (t.b.s.)	65.92	26.04	24.84
eu-es + BT (u.s.)	75.66	29.89	28.51
eu-es + BT (t.u.s.)	70.33	27.78	26.50
eu-es + BT (r.s.)	70.83	27.98	26.69
eu-es + BT (t.r.s.)	67.96	26.85	25.61
en-de	42.30	16.71	15.94
de-en	37.31	14.74	14.06
de-en + BT (b.s.)	51.53	20.35	19.42
de-en + BT (t.b.s.)	53.08	20.97	20.00
de-en + BT (u.s.)	54.37	21.48	20.49
de-en + BT (t.u.s.)	55.94	22.10	21.08
de-en + BT (r.s.)	52.26	20.64	19.69
de-en + BT (t.r.s.)	53.47	21.12	20.15
TOTAL			**355.53**

Table 8: Training time, power consumption and estimated CO_2 emissions for each system. 't.' stands for tagged; 'b.s.' for 'beam search'; 'u.s.' for 'unrestricted sampling'; and 'r.s.' for 'restricted sampling'.

a more efficient use of our GPUs, at the same time adjusting our estimation of the generated CO_2 emissions.

5. Ethical considerations and limitations

From the ongoing debate about ethical considerations and limitations of NLP systems [37], we identify at least 4 aspects that affect directly to the systems developed for the machine translation of clinical texts from Basque into Spanish:

1) Privacy: as mentioned in the end of the eu/es corpora description, both bilingual and monolingual clinical texts coming from hospitals were provided to us without any personally identifiable information (names, surnames, etc.), and

it was further de-identified by shuffling the sentences coming from each source. The authors had to sign a non-disclosure commitment before getting access to this private data.

2) Modality: since our system is limited to language in the form of text, cannot be applied to other uses based on speech. For instance, children that learn Basque as primary language cannot directly communicate with Spanish monolingual healthcare workers, forcing the parents to act as simultaneous (non professional) translators. The development of a speech translation system can overcome this issue, and could be considered as a possible future work once the system for translating text is implemented in Osakidetza. Specific solutions for people with non-verbal communication abilities could be addressed too, always in collaboration with professional interpreters.

3) Gender bias: in Section 4, the clinical terms 'patient', 'doctor' and 'nurse' have been defined in their most usual Spanish masculine and feminine forms, looking for their number of appearances in the compiled corpora and in the generated translations. This analysis will be extended to other appearances like '*al médico*' (masculine), the lack of diacritic in '*medico/a*', or the use of the less common '*doctor*' (masculine) and '*doctora*' (feminine). As future work, we plan to define a specific challenge test set for studying the gender bias in clinical domain MT from Basque into Spanish. Possible ways of correcting this bias will be discussed with the professionals from Osakidetza and implemented before system deployment.

4) Carbon footprint: in the end of Section 4, we have reported the training times for each system, along with the estimated power consumption and corresponding CO_2 emissions of our GPUs. In the future, we plan to extend this to consider also the power consumed by our CPUs, and automate the process by using some of the online available tools [38].[17] We will use this data to consider possible ways of reducing or neutralizing our carbon footprint.

[17] `https://github.com/Breakend/experiment-impact-tracker`

6. Conclusions and future work

In this work, we have analysed in detail the effect of BT on a real translation scenario as the clinical domain. For this purpose we have empirically compared and combined different methods for BT applied to the MT of clinical texts. One of the new combinations tried in this work, tagged restricted sampling, obtained the best automatic scores according to all the metrics studied in the eu-es direction. Unrestricted sampling obtained similar results, and has the advantage of not having to define another hyperparameter value as it happens with the top-k value in restricted sampling or the beam-width in beam search.

However, in the simulated low resource de-en scenario traditional beam search still obtained the best MT results, followed by tagged beam search. This confirms the generalized agreement that sampling is only helpful when large monolingual data are available. In any case, to drive more generalizable conclusions it would be necessary to try these methods on real low resource scenarios.

Considering also the LD metrics, the decoding algorithm that obtained the best MT results in the eu-es scenario (restricted sampling) obtained one of the lowest LD scores, so a human MT evaluation should be performed by bilingual healthcare workers to see which system actually provides the best translation quality. As a preliminary step, we have asked a bilingual biomedical expert to perform a human evaluation of the 3 systems that obtained higher MT evaluation scores, and observed that restricted sampling obtained significantly worse results than unrestricted sampling, even that the latter obtained lower MT automatic scores. This confirms our hypothesis that LD metrics can be used for complementing the MT automatic evaluation scores when identifying the best performing systems.

Analyzing our corpora/systems, we have detected gender-bias associated with socially stereotyped professions as doctor and nurse. For extending this analysis, we plan to design a contrastive test set for studying the gender bias in clinical domain MT from Basque into Spanish. With these results, the corresponding solutions to address this problem will be defined in collaboration with

Osakidetza before system deployment.

Finally, we have estimated the carbon footprint derived from our experiments. We will consider these values to study possible ways of reducing or neutralizing our carbon footprint.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) [grant number BES-2017-081045]; BBVA foundation's BigKnowledge project (2018); DOMINO project (MCIU / AEI / FEDER, UE) [grant number PGC2018-102041-B-I00]; and DOTT-HEALTH project (MCIU / AEI / FEDER, UE) [grant number PID2019-106942RB-C31].

References

[1] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, 2013, pp. 1700–1709.
URL https://www.aclweb.org/anthology/D13-1176

[2] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, in: Advances in Neural Information Processing Systems, Montréal, Canada, 2014, pp. 3104–3112.

[3] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proceedings of the 3rd International Conference on Learning Representations, San Diego, USA, 2015, 15pp.
URL http://arxiv.org/abs/1409.0473

[4] R. Sennrich, B. Haddow, A. Birch, 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), Berlin, Germany, 2016, pp. 86–96. doi:10.18653/v1/P16-1009.
URL https://www.aclweb.org/anthology/P16-1009

[5] C. Tillmann, H. Ney, Word reordering and a dynamic programming beam search algorithm for statistical machine translation, Computational Linguistics 29 (1) (2003) 97–133. doi:10.1162/089120103321337458.
URL https://www.aclweb.org/anthology/J03-1005

[6] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 489–500. doi:10.18653/v1/D18-1045.
URL https://www.aclweb.org/anthology/D18-1045

[7] I. Caswell, C. Chelba, D. Grangier, Tagged back-translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), Association for Computational Linguistics, Florence, Italy, 2019, pp. 53–63. doi:10.18653/v1/W19-5206.
URL https://www.aclweb.org/anthology/W19-5206

[8] M. Graça, Y. Kim, J. Schamper, S. Khadivi, H. Ney, Generalizing back-translation in neural machine translation, in: Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), Association for Computational Linguistics, Florence, Italy, 2019, pp. 45–52. doi:10.18653/v1/W19-5205.
URL https://www.aclweb.org/anthology/W19-5205

[9] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318.
URL https://www.aclweb.org/anthology/P02-1040

[10] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, USA, 2006, pp. 223–231.

[11] M. Popović, chrF: character n-gram f-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 2015, pp. 392–395. doi:10.18653/v1/W15-3049.
URL https://www.aclweb.org/anthology/W15-3049

[12] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, Michigan, 2005, pp. 65–72.
URL https://www.aclweb.org/anthology/W05-0909

[13] P. M. McCarthy, An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity, Ph.D. thesis, University of Memphis, TN (2005).

[14] M. C. Templin, Certain Language Skills in Children: Their Development and Interrelationships, University of Minnesota Press, Minneapolis, MN, 1975.

[15] G. U. Yule, The Statistical Study of Literary Vocabulary, Cambridge University Press, Cambridge, UK, 1944.

[16] F. Burlot, F. Yvon, Using monolingual data in neural machine translation: a systematic study, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Belgium, Brussels, 2018, pp. 144–155. doi:10.18653/v1/W18-6315.
URL https://www.aclweb.org/anthology/W18-6315

[17] X. Soto, O. Perez-De-Viñaspre, M. Oronoz, G. Labaka, Leveraging SNOMED CT terms and relations for machine translation of clinical texts from Basque to Spanish, in: Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation, Dublin, Ireland, 2019, pp. 8–18.
URL https://www.aclweb.org/anthology/W19-7102

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Long Beach, CA, 2017, pp. 5998–6008.

[19] A. Poncelas, M. Popovic, D. Shterionov, G. Maillette de Buy Wenniger, A. Way, Combining SMT and NMT Back-Translated Data for Efficient NMT, in: Proceedings of Recent Advances in Natural Language Processing, Varna, Bulgaria, 2019, pp. 922–931.

[20] X. Soto, D. Shterionov, A. Poncelas, A. Way, Selecting backtranslated data from multiple sources for improved neural machine translation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3898–3908. doi:10.18653/v1/2020.acl-main.359.
URL https://www.aclweb.org/anthology/2020.acl-main.359

[21] E. Biçici, D. Yuret, Optimizing instance selection for statistical machine translation with feature decay algorithms, Transactions on Audio, Speech & Language Processing 23 (2) (2015) 339–350.

[22] A. Poncelas, G. M. de Buy Wenniger, A. Way, Feature decay algorithms for neural machine translation, in: 21st Annual Conference of the European Association for Machine Translation, Alicante, Spain, 2018, pp. 239–248.

[23] B. Marie, R. Rubino, A. Fujita, Tagged back-translation revisited: Why does it really work?, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 5990–5997. doi:10.18653/v1/2020.acl-main.532.
URL https://www.aclweb.org/anthology/2020.acl-main.532

[24] E. Vanmassenhove, D. Shterionov, A. Way, Lost in translation: Loss and decay of linguistic richness in machine translation, in: Proceedings of Machine Translation Summit XVII (Research Track), Dublin, Ireland, 2019,

pp. 222–232.
URL https://www.aclweb.org/anthology/W19-6622

[25] E. Vanmassenhove, C. Hardmeier, A. Way, Getting gender right in neural machine translation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3003–3008. doi:10.18653/v1/D18-1334.
URL https://www.aclweb.org/anthology/D18-1334

[26] R. Sennrich, B. Haddow, A. Birch, 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), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. doi:10.18653/v1/P16-1162.
URL https://www.aclweb.org/anthology/P16-1162

[27] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, M. Auli, fairseq: A fast, extensible toolkit for sequence modeling, in: Proceedings of NAACL-HLT 2019: Demonstrations, 2019.

[28] I. Provilkov, D. Emelianenko, E. Voita, BPE-dropout: Simple and effective subword regularization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1882–1892. doi:10.18653/v1/2020.acl-main.170.
URL https://www.aclweb.org/anthology/2020.acl-main.170

[29] T. Etchegoyhen, A. Azpeitia, N. Pérez, Exploiting a large strongly comparable corpus, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), European Language Resources Association (ELRA), Portorož, Slovenia, 2016, pp. 3523–3529.
URL https://www.aclweb.org/anthology/L16-1560

[30] T. Etchegoyhen, H. Gete, Handle with care: A case study in comparable corpora exploitation for neural machine translation, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 3799–3807.
URL https://www.aclweb.org/anthology/2020.lrec-1.469

[31] I. San Vicente, I. Manterola, PaCo2: A fully automated tool for gathering parallel corpora from the web, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey, 2012, pp. 1–6.
URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/231_Paper.pdf

[32] I. Sarasola, P. Salaburu, J. Landa, Hizkuntzen Arteko Corpusa (HAC), University of the Basque Country UPV/EHU (Euskara Institutua), Bilbao, Spain, 2015.

[33] O. Perez-de-Viñaspre, Automatic medical term generation for a low-resource language: translation of SNOMED CT into Basque, Ph.D. thesis, University of the Basque Country, Donostia, Spain (2017).

[34] R. Bawden, G. M. Di Nunzio, C. Grozea, I. Jauregi Unanue, A. Jimeno Yepes, N. Mah, D. Martinez, A. Névéol, M. Neves, M. Oronoz, O. Perez-de Viñaspre, M. Piccardi, R. Roller, A. Siu, P. Thomas, F. Vezzani, M. Vicente Navarro, D. Wiemann, L. Yeganova, Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 660–687.
URL https://www.aclweb.org/anthology/2020.wmt-1.76

[35] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-

iii, a freely accessible critical care database, Scientific Data 3 (160035). doi:10.1038/sdata.2016.35.

[36] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. doi:10.18653/v1/P19-1355.
URL https://www.aclweb.org/anthology/P19-1355

[37] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. doi:10.1145/3442188.3445922.
URL https://doi.org/10.1145/3442188.3445922

[38] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, J. Pineau, Towards the systematic reporting of the energy and carbon footprints of machine learning, Computing Research Repository arXiv:2002.05651.
URL http://arxiv.org/abs/2002.05651