# Comparing and combining tagging with different decoding algorithms for back-translation in NMT: learnings from a low resource scenario

**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

Recently, diverse refinements to the back-translation process have been proposed for improving the performance of Neural Machine Translation (NMT) systems, including the use of sampling instead of beam search as decoding algorithm, or appending a tag to the back-translated corpus. However, not all the combinations of the previous approaches have been tested, remaining unclear which is the best approach for developing a given NMT system. In this work, we empirically compare and combine existing techniques for back-translation in a real low resource setting: the translation of clinical notes from Basque into Spanish. Apart from automatically evaluating the NMT systems, we ask bilingual healthcare workers to perform a human evaluation, and analyze the different synthetic corpora by measuring their lexical diversity. For reproducibility and generalizability, we repeat our experiments for German to English translation using public data. The results suggest that in lower resource scenarios tagging only helps when using sampling for decoding, complementing the previous literature using bigger corpora from the news domain. When fine-tuning with a few thousand bilingual in-domain sentences, one of our proposed methods (tagged restricted sampling) obtains the best results both in terms of automatic and human evaluation.



## 1 Introduction

Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) 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) (Sennrich et al., 2016a), 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 generating a translation, considering that looking for all the possible output sentences is practically infeasible, MT systems have to implement an efficient technique for selecting the most probable sentence according to the distribution of the training data. Typically, beam search (Tillmann and Ney, 2003) is used for generating both the output sentences of NMT systems and the synthetic sentences produced by BT systems.

Edunov et al. (2018) proposed to use sampling for BT as one way to further improve the performance of NMT systems. Specifically, their 'unrestricted sampling'[1] approach, consisting of randomly sampling from the output distribution, obtained the best results on average comparing to other decoding algorithms, including beam search.

On the contrary, Caswell et al. (2019) suggest that the improvement derived from using sampling

[1] In recent literature, unrestricted sampling is also referred as 'ancestral 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 Edunov et al. (2018), but they did not combine tagging with sampling.

Concurrent work by Graça et al. (2019) 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. From these combinations, the last two are proposed for the first time in this work:

1. beam search (Tillmann and Ney, 2003)
2. unrestricted sampling (Edunov et al., 2018)
3. restricted sampling (Graça et al., 2019)
4. tagged beam search (Caswell et al., 2019)
5. tagged unrestricted sampling (our contribution)
6. tagged restricted sampling (our contribution)

We compare these 6 methods both in terms of automatic evaluation of NMT systems, and lexical diversity (LD) of the synthetic corpora created by the BT systems. For MT automatic evaluation we use BLEU (Papineni et al., 2002), TER (Snover et al., 2006), chrF (Popović, 2015), and METEOR (Banerjee and Lavie, 2005); while for lexical diversity we measure TTR (Templin, 1975), Yule's I (Yule, 1944) and MTLD (McCarthy, 2005).

In the following, we briefly describe the lexical diversity metrics, for being less known.

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 typically used in MT.

As a complement to our MT and LD metrics, we add the results coming from a preliminary human evaluation done by a bilingual biomedical expert. According to these results, we select the best two systems for translating clinical reports from Basque to Spanish, and ask bilingual healthcare workers to post-edit the outputs of these systems, as well as the system trained in the opposite direction.

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 (Edunov et al., 2018; Graça et al., 2019; Caswell et al., 2019), 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 Burlot and Yvon (2018) firstly compared the use of statistical (SMT) and neural (NMT) systems for BT, without observing significant differences. More similarly to our work, Soto et al. (2019) tried rule-based (RBMT), SMT and NMT for BT ap-

plied to the translation of clinical texts, obtaining better results with NMT, and specifically the Transformer architecture (Vaswani et al., 2017).

Poncelas et al. (2019) 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, Soto et al. (2020) 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 (Biçici and Yuret, 2015; Poncelas et al., 2018) 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, Marie et al. (2020) 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 methods follows the work of Vanmassenhove et al. (2019), 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.

## 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.[2]

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, 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.[3] Further, we apply BPE (Sennrich et al., 2016b) for 90,000 (eu/es) and 40,000 (de/en) iterations. The number of BPE steps for eu/es was optimized in previous experiments, while the de/en one was taken from a reference system (Bawden et al., 2020) that will be described in Section 3.2.

For training all our systems, we use the Transformer architecture as implemented in Fairseq (Ott et al., 2019), with 6 encoder-decoder layers and an embedding size of 512.

All the systems were trained for 30 epochs, except the es–eu system that was trained for 50 epochs due to applying the BPE-dropout (Provilkov et al., 2020) regularization technique, 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 health records 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.

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.

[2]It is expected that the output of the MT system will be post-edited before making it available to Spanish monolingual healthcare workers.

[3]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

| 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 |
| health records | 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 5 million out-of-domain sentences. Among these, around 3 million sentences are from the news domain, formed by the 3 times repetition of a corpus from the Basque public broadcast service EiTB (Etchegoyhen et al., 2016), along with a more recent one from the same source (Etchegoyhen and Gete, 2020). The remaining 2 million sentences are from different domains as administrative (IVAP), consumer magazines (Eroski), online magazines (Irrika), translation memories (EIZIE), movie synopses, web crawling (San Vicente and Manterola, 2012) and literature (Sarasola et al., 2015).

We also 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 for this work we omitted the administrative ones (in Spanish: '*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 900,000 bilingual terms come from the automatic translation of SNOMED CT into Basque (Perez-de-Viñaspre, 2017), while another 30,000 are manual translations into Basque of ICD-10 concept descriptions in Spanish made available for the WMT Biomedical shared task (Bawden et al., 2020).

Finally, around 200 terms related to the COVID-19 pandemic are compiled, coming around half of them from an interim release of SNOMED CT that was made available in the beginning of the pandemic,[5] and translated into Basque by a translator of Osakidetza. The remaining terms were collected by Elhuyar.[6]

#### 3.1.3 Bilingual clinical notes

For fine-tuning and evaluation, we use the bilingual corpus compiled in the project with Osakidetza, where 149 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.

#### 3.1.4 Monolingual health records 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 2 million non-repeated sentences; as well as discharge (1 million) and progress (2 million) 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, 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 (Bawden et al., 2020), consisting of around 3 million sentences extracted from the UFAL cor-

[4] https://www.osakidetza.euskadi.eus/profesionales/-/publicaciones-profesionales/, accessed on October 1, 2020.

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

[6] We can make them available upon permission from Elhuyar.

pus[7] after removing the "Subtitles" subset. For evaluation we use Khresmoi,[8] also used in Bawden et al. (2020), 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 Bawden et al. (2020).

Finally, for BT we use the discharge reports in English available in MIMIC III (Johnson et al., 2016).[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 2 million sentences. We choose to not perform sentence splitting to avoid introducing errors associated with this process. As a consequence, 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 reproducing the de-en experiments can be found in `https://gitlab.com/xabiersoto1/bt_tagging_and_decoding`.

## 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).

[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/`

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

| | 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 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 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 Caswell et al. (2019), that presents tagged back-translation as a "simpler alternative to noising". With these results, we show that both tagging and sampling can be orthogonal methods to improve the performance 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,[12] 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.[13] Later we tokenize, truecase and apply BPE as done for the rest of the de/en data. Table 4 presents the results on HimL.[14]

| | 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 the hypothesis that unrestricted sampling only works with big corpora.

### 4.2 Lexical diversity 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.

| 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.

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 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, since the former will generally produce words that appear more in the training corpus. Considering these LD results, 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

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

[13] https://github.com/EdinburghNLP/nematus/blob/master/data/strip_sgml.py

[14] Specifically, on the 1044 sentences coming from the NHS subset, since the remaining sentences from Cochrane are used for validation purposes.

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 according to the most relevant MTLD metric, 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 as a proxy to assess the MT quality of different systems trained with the same corpus.

### 4.4 Human evaluation

In this section we present the results of the human evaluation performed by 37 bilingual healthcare workers. For doing this, we use PET[15] tool, asking each evaluator to post-edit 100 out of 500 sentences translated by the es–eu system and the best performing eu–es systems. Each of these 500 sentences was post-edited by 3 different evaluators. Considering that some of the sentences were translated equally by the two eu–es systems, 22 volunteers evaluated the eu–es translations, while 15 post-edited the outputs of the es–eu system.

Table 7 presents the post-editing times registered for each system. For a better comparison, we normalize the post-editing time by sentence length in the second column.

| | Seconds | Seconds/Word |
|---|---|---|
| es–eu | 65.88 | 7.19 |
| eu–es (tag. res. sam.) | 23.23 | 2.67 |
| eu–es (unr. sam.) | 22.78 | 2.66 |

**Table 7:** Average post-editing times by the best performing eu–es systems and the es–eu system, before and after normalizing per sentence length.

Comparing the results in each direction, we see that post-editing times are much larger for es–eu translation, while the difference between the two eu–es systems is very small, especially after normalization.

Table 8 shows the calculated HTER values, by distinguishing its post-edition types corresponding to insertions (INS), deletions (DEL), substitutions (SUB) and shifts (SHIFT).

| | HTER (ALL) | HTER (INS) | HTER (DEL) | HTER (SUB) | HTER (SHIFT) |
|---|---|---|---|---|---|
| es–eu | 12.47 | 0.95 | 3.39 | 7.21 | 0.92 |
| eu–es (t.r.s.) | 5.50 | 0.54 | 2.60 | 2.17 | 0.20 |
| eu–es (u.s.) | 6.24 | 0.60 | 3.00 | 2.30 | 0.35 |

**Table 8:** HTER values by the best performing eu–es systems and the es–eu system, disaggregated by post-edition types.

As it happened with post-editing times, we observe that the HTER values are higher for the es–eu direction. On the other hand, while post-editing times were slightly higher for the 'tagged restricted sampling' system, we see that this system outperforms the 'unrestricted sampling' system regarding HTER and all its post-edition types.

Finally, Table 9 shows the average keystrokes registered by PET in all its 3 main values.

| | VISIBLE | KEYSTROKES | ALLKEYS |
|---|---|---|---|
| es–eu | 7.32 | 10.20 | 11.13 |
| eu–es (t.r.s.) | 3.23 | 4.21 | 4.42 |
| eu–es (u.s.) | 4.16 | 5.41 | 5.63 |

**Table 9:** Registered keystrokes for the best performing eu–es systems and the es–eu system, where "VISIBLE": letters + digits + spaces + symbols; "KEYSTROKES": "VISIBLE" + erase; and "ALLKEYS": "KEYSTROKES" + navigation + commands.

Again, for the eu–es direction, we see that the 'tagged restricted sampling' system obtains better results than the 'unrestricted sampling' system in

[15] https://github.com/wilkeraziz/PET

terms of keystrokes, so we select this system for a final error analysis.

### 4.5 Error analysis

Table 10 shows the number of omissions, additions, mistranslations and shift errors by the best performing 'tagged restricted sampling' system in the eu–es direction, distinguishing between single and multiple word errors.

| | Omissions | Additions | Mistransl. | Shifts |
|---|---|---|---|---|
| TOTAL | 51 | 6 | 103 | 4 |
| Single words | 35 | 4 | 80 | 1 |
| Multiple words | 16 | 2 | 23 | 3 |

**Table 10:** Classification of the MT errors for the best performing eu–es system (tagged restricted sampling).

We observe that most of the errors correspond to mistranslations, approximately doubling the omissions, and being the additions and shifts very scarce. For the most common omissions and mistranslations, most of the time these errors are related to a single word, especially for the latter.

From the omitted words, 12 are articles, while one of the added words is also an article. Among the mistranslations, there are 15 clinical terms translated as acronyms, 8 abbreviations, 3 missing accents and 3 singular/plural mismatches. Notice that all of these errors will not substantially alter the sentence meaning.

### 4.6 Carbon footprint

To conclude this section, answering to the call made by Strubell et al. (2019), 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 11 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 Strubell et al. (2019), 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 a more efficient use of our GPUs, at the same time adjusting our estimation of the generated $CO_2$ emissions.

| System | Time (h) | Power (kWh) | $CO_2$e (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 11:** 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'.

## 5 Conclusions and future work

In this work, 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, confirmed by the HTER and keystroke results from the human evaluation performed by bilingual healthcare workers.

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. Moreover, we observe that tagging only helps when using sampling for decoding the BT systems, complementing previous work that proposed tagging the synthetic corpora as an alternative to the use of sampling. However, to drive more generalizable conclusions it would be necessary to try these methods on more diverse scenarios.

Considering the LD metrics, the decoding algorithm that obtained the best MT results in the eu–es scenario (restricted sampling) obtained one of the lowest MTLD scores. In a preliminary human

evaluation done by a bilingual biomedical expert to assess the 3 systems that obtained higher MT evaluation scores, restricted sampling obtained significantly worse results than unrestricted sampling, even that the latter obtained lower MT automatic scores. This is a sign that LD metrics can be used as a complement to the MT automatic evaluation scores for identifying the best performing systems.

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

We thank the healthcare workers who volunteered compiling the bilingual clinical domain corpus and taking part in the human evaluation. We also thank Nora Aranberri and Ekain Arrieta for helping us with the human evaluation, as well as Marco Turchi and Luisa Bentivogli for fruitful discussions. This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) [grant number BES-2017-081045]; DOTT-HEALTH project (MCIU / AEI / FEDER, UE) [grant number PID2019-106942RB-C31]; and both by the Spanish Ministry of Science and Innovation and the European Commission in a CHIST-ERA project (FEDER, ANTIDOTE PCI2020-120717-2).

## References

Banerjee, Satanjeev, 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*, Ann Arbor, Michigan, USA. 65–72.

Bawden, Rachel, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de-Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 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*, online. 660–687.

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

Burlot, Franck, 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*, Brussels, Belgium. 144–155.

Caswell, Isaac, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy. 53–63.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. 489–500.

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*, Porotoroz, Slovenia. 3523–3529.

Etchegoyhen, Thierry, and Harritxu 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*, Marseille, France. 3799–3807.

Graça, Miguel, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. 2019. Generalizing back-translation in neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, Florence, Italy. 45–52.

Johnson, Alistair E.W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(160035).

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

Marie, Benjamin, Raphael Rubino, and Atsushi Fujita 2020. Tagged Back-translation Revisited: Why Does It Really Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, online. 5990–5997.

McCarthy, Philip M. 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, Tennessee, USA.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Minneapolis, Minnesota, USA. 48–53.

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*, Philadelphia, Pennsylvania, USA. 311–318.

Perez-de-Viñaspre, Olatz. 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.

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

Poncelas, Alberto, Maja Popović, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Combining PBSMT and NMT Back-translated Data for Efficient NMT. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria. 922–931.

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

Provilkov, Ivan, 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*, online. 1882–1892.

San Vicente, Iñaki, and Iker Manterola. 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*, Istanbul, Turkey. 1–6.

Sarasola, Ibon, Pello Salaburu, and Josu Landa. 2015. *Hizkuntzen Arteko Corpusa (HAC)*. University of the Basque Country UPV/EHU (Euskara Institutua), Bilbo, Spain.

Sennrich, Rico, 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)*, Berlin, Germany. 86–96.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany. 1715–1725.

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

Soto, Xabier, 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*, Dublin, Ireland. 8–18.

Soto, Xabier, Dimitar Shterionov, Alberto Poncelas, and Andy Way. 2020. Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, online. 3898–3908.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. 3645–3650.

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

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

Tillmann, Christoph, and Hermann Ney. 2003. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational Linguistics*, 29(1):97–133.

Vanmassenhove, Eva, 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*, Dublin, Ireland. 222–232.

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*, Long Beach, California, USA. 5998–6008.

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