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) 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):

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

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):

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

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		+RBMT$_{bt}$	23.27	62.67	48.02	56.51
	Auth.	+PB-SMT$_{bt}$	22.51	64.57	45.97	54.53
	+	+LSTM$_{bt}$	24.74	63.55	47.58	55.59
	BT.	+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***
	+	EachFromAll x4	24.59	61.15	48.10	56.19
	DS	EachFromAll RS	25.77	59.86	48.59	56.92
DE-EN		+RBMT$_{bt}$	39.02	42.27	37.32	62.72
	Auth.	+PB-SMT$_{bt}$	42.32	39.21	39.37	65.91
	+	+LSTM$_{bt}$	40.97	39.75	38.45	64.81
	BT	+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.

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.

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

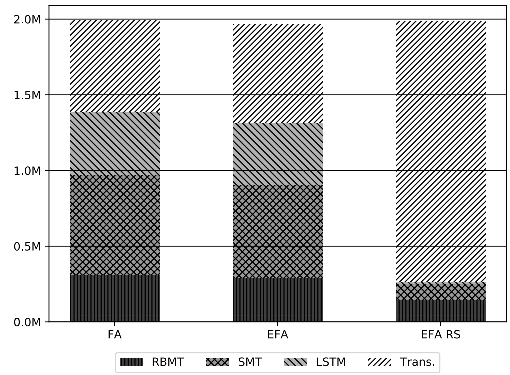

Figure 1: Amount of sentences selected from each system by the data selection approaches for EU-ES. FA = FromAll, EFA = EachFromAll, EFA RS = EachFromAll Rescored.

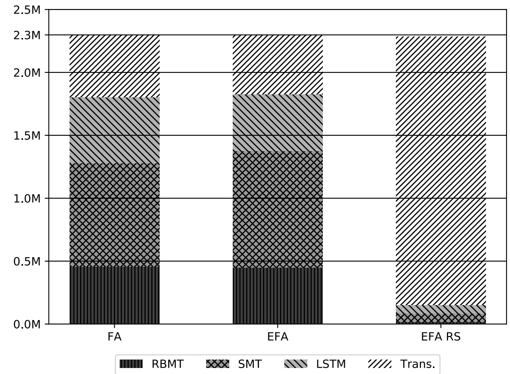

Figure 2: Amount of sentences selected from each system by the data selection approaches for EN-DE. FA = FromAll, EFA = EachFromAll, EFA RS = EachFromAll Rescored.

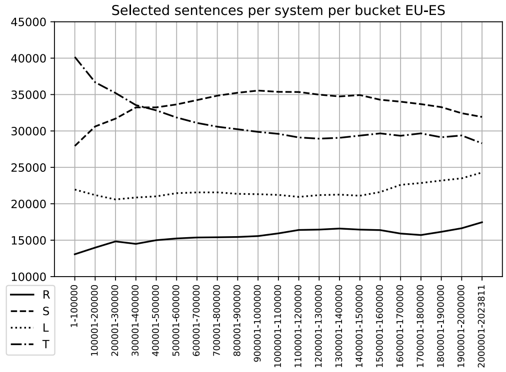

Figure 3: Number of sentences selected from each system by the FromAll data selection approach for EU-ES language pair in subsequent bins of 100,000 sentences (extrapolated for the last bin).

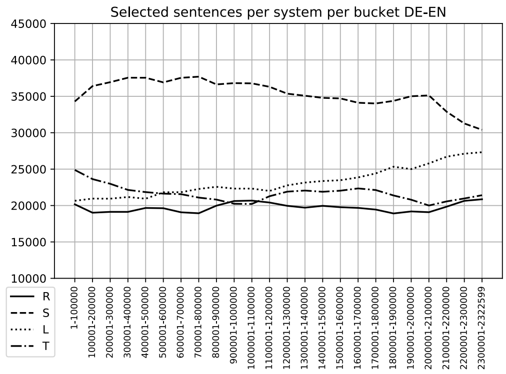

Figure 4: Number of sentences selected from each system by the FromAll data selection approach for DE-EN language pair in subsequent bins of 100,000 sentences (extrapolated for the last bin).

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.