This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

Health in my Language

HimL

D1.2: Report on Improving Translation with Monolingual Data

Author(s): Fabienne Braune, Alex Fraser, Barry Haddow

Dissemination Level: Public

Date: January, 31st 2018

Version 1.0

Grant agreement no.	644402
Project acronym	HimL
Project full title	Health in my Language
Funding Scheme	Innovation Action
Coordinator	Barry Haddow (UEDIN)
Start date, duration	1 February 2015, 36 months
Distribution	Public
Contractual date of delivery	January, 31st 2018
Actual date of delivery	January, 31st 2018
Deliverable number	D1.2
Deliverable title	Report on Improving Translation with Monolingual Data
Type	Report
Status and version	1.0
Number of pages	15
Contributing partners	UEDIN, LMU
WP leader	UEDIN
Task leader	LMU
Authors	Fabienne Braune, Alex Fraser, Barry Haddow
EC project officer	Tünde Turbucz
The Partners in HimL are:	The University of Edinburgh (UEDIN), United Kingdom
	Univerzita Karlova V Praze (CUNI), Czech Republic
	Ludwig-Maximilians-Universitaet Muenchen (LMU-MUENCHEN), Germany
	Lingea SRO (LINGEA), Czech Republic
	NHS 24 (Scotland) (NHS24), United Kingdom
	Cochrane (COCHRANE), United Kingdom

For copies or reports, updates on project activities and other HimL-related information, contact:

Barry Haddow bhaddow@staffmail.ed.ac.uk
University of Edinburgh Phone: +44 (0) 131 651 3173

Contents

1 Introduction

In this report we present the results of research into the use of monolingual data to improve machine translation systems. There are really two areas of investigation pursued here; bilingual lexicon induction from comparable corpora, and using monolingual data in neural MT (mainly through back-translation to create synthetic data). The first of these areas was the one envisaged in Task T1.3 of the Description of Action, however the second became very relevant as the project progressed, and NMT took over as the state-of-the-art paradigm in MT.

Before we present the results of bilingual lexicon induction in Section 3, we will first present results of a training corpus coverage study, which measures the extent of the out-of-vocabulary problem on the HimL use-cases.

2 Coverage Analysis

In this section we investigate how the accuracy of translation for domain-specific terms is affected by training set coverage. This is an extension of the analysis of *D1.1: Report on Building Translation Systems for Public Health Domain*, where we tried to link domain adaptation techniques with translation accuracy of domain-specific terms. The analysis here is applied to phrase-based MT, rather than NMT as is used in Y3 of HimL, because we build on this earlier work, but we do not expect substantially different results for NMT since it uses the same training data. We proceed by first identifying terms in the HimL 2015 test sets, then defining and calculating measures of domain-specificity, coverage and translation accuracy for these terms.

The term extraction, domain-specificity and translation accuracy measurements follow D1.1, and are described fully in Section 2.3.3 there. Term extraction is based on the chunks created by TreeTagger.[1] For the current analysis, we also add a coverage calculation and extraction of the reference translation:

Calculation of Coverage We define the coverage of a (potentially multi-word) term by considering all possible *partitions* of the term. A partition of a given string of words is a splitting into separate, non-overalpping substrings, for example "randomised controlled trial" can be partitioned into "randomised controlled" and "trial"; or "randomised", "controlled" and "trial"; as well as several other possible partitions. We define the coverage of a partition as the minimum number of occurrences in the training data over each segment in the partition. We then define the coverage of a term as the maximum coverage over all possible partitions. This definition captures the intuition that a multi-word term should still be translated correctly even if it is rare or unseen in training, if it can be decomposed into well-covered segments (although it ignores the fact that many terms cannot be translated compositionally).

Reference Translation of Terms We extract reference translations of all terms using a word alignment of the test set. The word alignment is created by concatenating the test set onto the training set and running `fast_align` (Dyer *et al.*, 2013), and then the translation of a source term can be read off by projecting through the alignment.

We calculated the translation accuracy using the "provenance" model of D1.2, with the best-performing language model combination. For each of the 4 target languages of HimL (Czech, German, Polish and Romanian) and each of the 2 domains (NHS24 and Cochrane) we select the terms which are being consistently incorrectly translated (i.e. they have accuracy zero). We restrict our attention to terms with a domain specificity measure of greater than 1, where the measure is the difference between the normalised log probability of in-domain and out-of-domain language models. We then rank the terms by training set coverage (as defined above) from lowest to highest, and show the selected terms with lowest coverage in Appendix A. We show the first page of terms with the lowest coverage, for each domain-language combination.

Examining the tables in Appendix A, we see that for the NHS24 test sets, few of the badly translated terms have poor coverage in the parallel data, whereas for Cochrane most of the terms shown in the tables have coverage of less than 10. This indicates that the Cochrane data tends to contain more specialist, domain-specific language, and that such language causes a problem for MT systems. NHS24 texts aim to employ simple English where possible. The tables show a mixture of proper names (where the initials and surnames are sometimes transposed in the reference), unusual capitalisation, and genuine low coverage terms. It should be noted that a badly translated term can impact the whole sentence as the MT system does not know how to order the sentence correctly, and neighbouring words lack target-side evidence.

Examining the terms in the appendix in more detail, we note some trends. We can see that genuine OOVs do occur (e.g. "subfertile" and "thromboelastometry") but sometimes (e.g. in the latter term) have quite similar translations in the target language anyway. We can observe the following types of issues in the translation:

- Numbers, times and names appear fairly frequently in the low-coverage term list, and are not always well translated. These could potentially be fixed with rule-based processing coupled with place-holders.

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

- There are cases of incorrect morphology on translation of low-coverage terms, for instance the translations of “okra” and “balloon” in the en-cs NHS24 set.
- Incorrect pass-throughs[2] (e.g. “Peto odds ratio”), although this is not strictly a pass-through as “odds” and “ratio” are not OOVs.
- Inconsistent or incorrect translation of very specific terminology (e.g. “TEF”, “Rotem” and “RCT” in Cochrane data). Sometimes it’s just a capitalisation issue.
- The limitations of the analysis method are exposed on discontinuous translations such as coordinations, where the reference translations cannot easily be found by projecting through the alignment. For example in NHS24 en-pl. The source is "Monday to Friday 9am to 5pm", and the hypothesis is "od poniedziałku do piątku od 9 do 17" but ref is "od poniedziałku do piątku między 9 a 17". The problematic term is identified as "Friday 9am", with hypothesis as "piątku 9", marked wrong since the hypothesis is not found in the reference.

We conclude that, even though we use large (50 million or more sentences) training sets for HimL, there are still gaps in the coverage that need to be filled by exploiting other sources of data.

3 Bilingual Lexicon Induction

Bilingual lexicon induction (BLI) is the task of generating, given a list of words in a source language, accurate translations for each word. The possibility to perform BLI *without parallel data* is critical in many low resource scenarios. Bilingual word embeddings (BWEs), where words from different languages are represented in the same vector space, have shown very effective to perform BLI given a *small* seed lexicon (5000 word-pairs) as the only bilingual signal, see e.g. Mikolov *et al.* (2013b); Faruqui and Dyer (2014); Lazaridou *et al.* (2015); Xing *et al.* (2015); Vulic and Korhonen (2016). Bilingual lexicons extracted using BWEs have been evaluated on frequent words from parliament proceedings or Wikipedia articles and shown good accuracies on these datasets. However, evaluations on rare and domain-specific words have not yet been provided although such evaluation scenarios are critical for applications like machine translation or bilingual terminology mining.

To address this, we design a novel evaluation scenario for BWEs: given (i) large amounts of monolingual data and a (ii) small seed lexicon of frequent word-pairs, the goal is to create BWEs that enable accurate mining of rare words. We consider two types of rare words: (i) very low occurrence frequencies (3 to 5) in different domains; and (ii) medical terms. Since (i) is important in many machine translation scenarios, we show extensive experiments on these types of words. We also focus on (ii) since it is particularly relevant for HimL.

We begin by showing that, on rare words, commonly used approaches to BWEs perform poorly. We present simple ways to build and combine BWEs that yield large performance improvements over previous work and constitute strong baselines for BLI of rare words-pairs. Finally, we show that our techniques are not only useful for rare and domain specific BLI but also yields performance improvements over state-of-the-art approaches on commonly used evaluation scenarios (on Wikipedia articles or parliament proceedings). We make training and test as well as baselines for our task publicly available for further research.

3.1 Bilingual induction of rare and in-domain words

Our training set for BWEs consists of two large corpora. Although we work with parallel data, we use it in a monolingual way by shuffling each side of the parallel corpus. We experimented with true monolingual data in the general domain and noted small decreases in performance over shuffled parallel.

- GENERAL: 4,400,309 English and German sentences from parliament proceedings, news commentaries and web crawls taken from the WMT 2016 shared task.
- MEDICAL: 3,108,183 English and German sentences from of titles of medical Wikipedia articles, medical term-pairs, patents, documents from the European Medicines Agency. This is the in-domain part of the UFAL corpus.[3]

As seed lexicon, we take the 5000 most common words in GENERAL and MEDICAL and translate those using a probabilistic dictionary.[4] BWEs trained using this data are evaluated on two gold standards containing pairs of rare words.

[2] A pass-through is a term which is OOV to the translation system, so it passes it through unchanged to the hypothesis. This only happens with statistical MT, not with neural MT

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

[4] This dictionary is taken from a standard English/German phrase-based system built on WMT 2017 data.

Low frequency word-pairs This gold standard is created by randomly sampling words occurring between 3 and 5 times[5] in GENERAL and MEDICAL. For GENERAL we sample rare words from news commentaries and web crawls separately. For each (English) sampled word, a German native speaker generated a German translation. Gold standard data is divided into validation and test sets as follows:

- CRAWLR: 1000 rare words from web crawls (250 validataion, 750 test)
- NEWSR: 1169 rare words from news texts (369 validation, 800 test)
- MEDR: 2000 rare words form medical texts (1000 validation, 1000 test)

Medical word-pairs Besides very low frequency terms, we create a gold standard of word-pairs in the medical domain. We sample the 2000 most frequent words after the seed lexicon form MEDICAL and automatically translate these using a probabilistic dictionary. A random sample of 1000 words are used for validation and 1000 other words are used for test. Note that there is previous work on BLI of medical terms. BLI of English-Dutch medical terms has been addressed by Heyman *et al.* (2017) who work in a scenario where very small amounts of document-aligned medical texts are available. Our task is different from theirs in that we work with large amounts of monolingual medical data without requiring any aligned documents.

3.2 Bilingual word embedding creation

To create bilingual word embeddings, we use *post-hoc mapping* (PHM), a method that projects monolingual word embeddings into a shared space using a linear transformation trained with a small seed lexicon, see Mikolov *et al.* (2013b); Faruqui and Dyer (2014); Xing *et al.* (2015); Lazaridou *et al.* (2015); Vulic and Korhonen (2016). Among methods to generate BWEs, PHM uses the cheapest bilingual signal, just a seed lexicon.[6]
Given monolingual word embeddings in two languages V_s and V_t, the goal of post-hoc mapping is to find a matrix $\mathbf{W} \in \mathbb{R}^{d_1 \times d_2}$ that maps each representation $\vec{x_i} \in \mathbb{R}^{d_1}$ of a source word $s \in V_s$ to the representation $\vec{y_i} \in \mathbb{R}^{d_2}$ of its translation $t \in V_t$. Typically, $\mathbf{W}$ is learned using a seed lexicon $L = \{(\vec{x_1}, \vec{y_1}), \ldots, (\vec{x_n}, \vec{y_n})\}$, where each pair $(\vec{x_i}, \vec{y_i})$ represents words in V_s and V_t that are mutual translations. A common objective to cross-lingual mapping is ridge regression (Mikolov *et al.*, 2013b) (RIDGE), where $\mathbf{W}$ is estimated by:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathbb{R}^{\mathbf{d_1} \times \mathbf{d_2}}}{\arg\min} \parallel \mathbf{XW} - \mathbf{Y} \parallel + \lambda \parallel \mathbf{W} \parallel \quad (1)$$

Lazaridou *et al.* (2015) use a max-margin ranking loss (MAX-MARG) to estimate $\mathbf{W}$. For each representation $\vec{x_i}$ of a word $s \in V_s$ in L, a candidate $\vec{y^*} = \mathbf{W} \cdot \vec{x_i}$ is computed. The ranking loss is then given by:

$$\sum_{i \neq j}^{k} \max\{0, \gamma + S\,dist(\vec{y^*}, \vec{y_i}) - S\,dist(\vec{y^*}, \vec{y_j})\} \quad (2)$$

where $\vec{y_i}$ corresponds to $\vec{x_i}$ in L. The function $S\,dist(\vec{x}, \vec{y})$ computes the semantic distance between $\vec{x}$ and $\vec{y}$ using inverse cosine. The hyperparameter γ is tuned on held-out validation data.[7]
We reimplement Mikolov *et al.* (2013b) as well as Lazaridou *et al.* (2015). To replicate their results on English-German texts[8], we evaluate these on GENERAL. First, we train monolingual word embeddings on the monolingual data using w2v (Mikolov *et al.*, 2013a) skip-gram (w2v SKIP) and cbow (w2v CBOW). o generate BWEs, we apply RIDGE and MAX-MARG using a seed lexicon with the 5000 most frequent English words from GENERAL. The results in Table 1 show that training monolingual word embeddings with w2v CBOW and mapping with RIDGE yields the best results on our dataset. Accuracies are comparable to previous work on different language pairs. As in previous work, we report top-1 (translation is closest neighbor) and top-5 (translation is one of 5 closest neighbors) accuracies.

3.3 Applying BWEs to rare word-pairs

We use the exact same BWEs training setup as above (3.2) and perform BLI on our test sets of rare words. The results in Table 2 show that on **low frequency** word-pairs BWEs perform very poorly. Compared to standard evaluation scenarios (see Table 1)

[5] Words with frequencies 1 and 2 are very often tokenization errors or borrowings from other languages, therefore we start at frequency 3. We did not consider tokenization errors as rare words and removed those from our data.
[6] Gouws and Søgaard (2015); Duong *et al.* (2016) also leverage seed lexicons. However, in order to generate high quality BWEs, these approaches leverage much larger bilingual dictionaries.
[7] Ideally, the sum in Equation 2 should be computed over the complete target vocabulary (i.e. $k = |V_t|$). Since this is not feasible in practice, Lazaridou *et al.* (2015) treat k as another hyperparameter tuned together with γ.
[8] These approaches have not yet been evaluated on English-German texts. Before applying on our task, we make sure that we obtain comparable results on frequent and general domain words.

Domain	mapping	w2v skip	w2v cbow
General	ridge	32.5 (43.4)	**40.9 (61.0)**
General	max-marg	31.6 (47.1)	27.9 (45.7)

Table 1: Bilingual lexicon induction of frequent word-pairs on general domain data. We report top-1 and top-5 accuracies (top-5 in brackets).

a massive performance decrease is observed. The low accuracy is most likely caused by the inability of context-based models (w2v) to build accurate embedding vectors for words occurring in very few contexts only. Through post-hoc mapping, these (poor) embeddings get projected randomly into the bilingual space which results in very poor performance on BLI. Mining (frequent) **medical word-pairs** yields better performance (shown in Table 3) but BLI accuracy is still low compared to results obtained on general domain data (Table 1).

Domain	mapping	w2v skip	w2v cbow
crawlRare	ridge	2.2 (3.2)	2.0 (2.4)
crawlRare	max-arg	2.4 (3.1)	1.8 (2.3)
newsRare	ridge	4.6 (9.4)	2.1 (5.3)
newsRare	max-marg	5.5 (11.0)	2.1 (4.9)
medRare	ridge	1.6 (2.8)	1.4 (2.3)
medRare	max-marg	1.8 (3.6)	1.3 (2.5)

Table 2: Bilingual lexicon induction of low frequency word-pairs in different domains.

Domain	mapping	w2v skip	w2v cbow
Medical	ridge	12.4 (22.2)	**16.6 (30)**
Medical	max-marg	15.7 (26.7)	15.3 (28.4)

Table 3: Bilingual lexicon induction of (frequent) medical words-pairs.

3.3.1 Using subword models

In order to create BWEs that are better adapted to rare and domain-specific words, we try to generate monolingual word embeddings that provide better vector representations for these words. We apply FastText (Bojanowski *et al.*, 2016), which changes w2v by using subword information $s(w, c)$ as the context-based objective as follows:

$$s(w, c) = \sum_{g \in G_w} z_g^\top v_c$$

where $G_w \subset \{1, ..., G\}$ is the set of n-grams that appear in the word w, z_g is the vector representation of the n-gram g and v_c is the vector of the context words. Subword information may alleviate the lack of context available for rare words and generate more accurate monolingual word embeddings. We create monolingual word embeddings using Fasttext skip-gram and cbow models with default parameters. We perform PHM using ridge and max-margin. The results in Table 4 show that this procedure yields impressive performance improvements on all datasets. Generating BWEs with max-margin on these improved monolingual word embeddings is particularly effective.

While impressive improvements are observed on our task (e.g., 30.7 top-1 score in Table 5), monolingual word embeddings trained with fasttext lead to a small performance decrease on general domain data, shown in Table 6.

3.3.2 Model Ensembling

Although BWEs obtained with fasttext and max-margin clearly outperforms other methods on rare words, a combination of BWEs obtained with different models may further improve performance by integrating several sources of information. Additionally, it may work well on rare words without causing a performance drop on general domain data. We ensemble BWEs obtained using different monolingual word embeddings as follows: we generate n-best lists of translation candidates using each

Domain	mapping	FTT skip	FTT cbow
crawlRare	ridge	10 (14)	6.7 (10.8)
crawlRare	max-marg	**11.7 (16.4)**	7.3 (12.8)
newsRare	ridge	23.2 (37.6)	6.7 (13.5)
newsRare	max-marg	**26.4 (40.1)**	14.9 (23.5)
medRare	ridge	12.1 (19.02)	7.2 (13.4)
medRare	max-marg	**12.3 (20.1)**	8.5 (15.6)

Table 4: Bilingual lexicon induction of low frequency word-pairs using ensembles of BWEs. monolingual word embeddings are trained with FASTTEXT (FTT).

Domain	mapping	FTT skip	FTT cbow
Medical	ridge	20.5 (35.3)	16.0 (28.1)
Medical	max-marg	**30.7 (43.4)**	23.9 (37.2)

Table 5: Bilingual lexicon induction of medical word-pairs. monolingual word embeddings are trained with FASTTEXT (FTT).

Domain	mapping	FTT skip	FTT cbow
General	ridge	32.1 (52.2)	16.9 (32.2)
General	max-marg	38.9 (56.8)	27.9 (45.7)

Table 6: Bilingual lexicon induction of frequent word-pairs in the general domain.

model. For each pair (s, t) of candidate translations, we compute an ensemble weight given by a weighted sum of similarity scores $\text{Sim}_i(s, t)$ obtained on each BWE:

$$\sum_{i=1}^{M} \gamma_1 \text{Sim}_i(s, t) + \ldots + \gamma_M \text{Sim}_M(s, t) \tag{3}$$

$\text{Sim}_i(s, t)$ is computed using cosine similarity. When a candidate pair (s, t) is not in the list generated by a model[9] i then $\text{Sim}_i(s, t)$ is set to 0. The weights γ_i are tuned on validation sets for our task (presented in 3.1), using grid search. The results, displayed in Tables 7 and 8, show that ensembling yields small gains over subword models. however, while subword models decreased performance on GENERAL ensembling also boosts performance on this dataset. The results are shown in Table 9. For **medical** word-pairs, orthographic distance further boosts performance but without massive gains. Finally, our method also yields improvements on general domain data.

Domain	mapping	ensemble BWEs
crawlRare	ridge	10.3 (14.5)
crawlRare	max-marg	13.5 (17.5)
newsRare	ridge	25.3 (40.0)
newsRare	max-marg	**27.74 (40.7)**
medRare	ridge	**15.1** (20.6)
medRare	max-marg	**14.2 (21.2)**

Table 7: Bilingual lexicon induction of low frequency word-pairs using ensembles of BWEs. Ensemble BWEs denotes an ensemble of four BWEs obtained with w2v and FASTTEXT (skip-gram and cbow) and mapped with RIDGE and MAX-MARGIN.

Instead of ensembling similarity scores we tried to work with richer monolingual word embeddings obtained with average and concatenation of embedding vectors obtained with w2v and FASTTEXT. We then applied RIDGE and MAX-MARG on these. The obtained BWEs did not yield performance gains on our tasks. We conjecture that we were not able to learn an accurate linear

[9] Because the models generate different n-best lists, certain word pairs may be generated by a model but not by the others.

Domain	mapping	ensemble BWEs
Medical	ridge	21.13 (37.5)
Medical	max-marg	**32.7 (44.1)**

Table 8: Bilingual lexicon induction of medical word-pairs using ensembles.

Domain	mapping	ensemble BWEs
General	ridge	**45.1 (67.4)**
General	max-marg	31.2 (46.1)

Table 9: Bilingual lexicon induction of frequent word-pairs in the general domain using ensembles.

mapping (using our small seed lexicon) for these richer representations because the richer representation makes it easier for us to overfit.

3.3.3 Adding orthographic distance

While subword information captures orthographic properties of words to a certain extent, it may be beneficial to strengthen BWEs by integrating a similarity measure between word surface forms only. The BWEs ensemble in Equation 3 can easily be augmented with a weighted term $\gamma_{M+1}\text{Odist}(s,t)$ that measures the orthographic distance (normalized Levensthein distance) between the surface-forms of words s and t.[10] We generate n-best lists of candidate translations using different BWEs models as in 3.3.2. In addition, we generate a list containing the n closest target words according to $\text{Odist}(s,t)$ and ensemble all lists together. The results are shown in Tables 10 and 11. For conciseness, we only report the results obtained when adding orthographic information to our best performing ensembles in 3.3.2. The results show that for **low frequency** word-pairs, orthographic information leads to massive performance gains. To measure the impact of orthographic information only, we also report the results obtained when using this information only (all other ensemble weights set to 0). On **medical** word-pairs, orthographic information further improves performance. Finally, our technique also yields performance gains on general domain data.

Domain	mapping	ensemble + edit	edit only
crawlRare	max-arg	**25.8 (29.01)**	24.8 (28.85)
newsRare	max-marg	**30.0 (41.29)**	20.51 (25.67)
medRare	max-marg	**28.0 (30.3)**	27.8 (29.9)

Table 10: Bilingual lexicon induction of low-frequency word-pairs in different domains. "Edit only" denotes the results obtained by using only orthographic distance (all other weights set to 0).

Domain	mapping	ensemble + edit	edit only
Medical	max-marg	**34.2 (45.0)**	21.6 (33.9)
General	ridge	**47.1 (63.9)**	16.5 (27.2)

Table 11: Bilingual lexicon induction of frequent words in different domains, only the best-performing ensembles are shown

3.4 Conclusion

We have evaluated BWEs for BLI in scenarios that are particularly useful for domain-specific machine translation and bilingual terminology mining. We have shown that state-of-the art approaches fail to perform well in these scenarios. By ensembling different BWEs and combining those with orthographic cues, we have massively improved BLI for such scenarios and hence provided strong baselines for BLI on rare and domain-specific terms. By making our code and datasets publicly available, we encourage further work on enhancing BWEs for these tasks.

[10] We experiment with a very simple orthographic measure but $\text{Odist}(s,t)$ could be computed using a more sophisticated model.

4 Monolingual Data in Neural MT

In this section we describe further experiments on the use of monolingual data in neural machine translation (NMT). In earlier statistical systems, large language models were the norm, often built from enormous quantities of monolingual data (e.g. Durrani *et al.* (2014)). In NMT, it was initially unclear how to incorporate monolingual data, but the most commonly used route today is *back-translation* (Sennrich *et al.*, 2016a)[11]. This means that we use monolingual target text to create synthetic parallel text by automatically translating it into the source language (using an NMT system for the reverse direction). The synthetic, back-translated data can be combined with naturally occurring parallel data in NMT training, or can be used for continued training of an already converged model.

We showed some results on using back-translated data to improve HimL systems in *D1.1: Report on Building Translation Systems for Public Health Domain*, and here we offer a much more extensive set of experiments.

4.1 Selecting and using back-translated monolingual data

The basic idea for using back-translated monolingual data in HimL was described in *D1.1*. Since we want the back-translated data to help with domain adaptation, but do not have any clearly in-domain target language data, we use in-domain source language data, automatically translated into the target language, to select from CommonCrawl (Buck *et al.*, 2014). The in-domain source language data is gathered from the Cochrane and NHS 24 websites, and after translating it we use Moore-Lewis selection to find appropriate target language sentences. The synthetic data is combined with parallel data from the EMEA (European medicine agency) and small amounts of Cochrane translations (about 10k sentences for en-de and 1k for en-pl) to create the in-domain corpus. The in-domain corpus is mixed 1:1 with the general parallel corpus, either in a separate fine-tuning step, or from the beginning of training.

We used this technique – selection from CommonCrawl and back-translation – in our Y3 systems (see *D4.3-6: Deployed translation systems*) and in Edinburgh's submissions to the WMT17 biomedical translation task (Sennrich *et al.*, 2017). In general it worked well, providing gains of up to 4.7 BLEU over baseline systems. Of the four HimL language pairs, the only one where the back-translation technique did not offer consistent improvements was en-ro. The problem here was that much of the Romanian CommonCrawl text uses diacritics inconsistently (sometimes dropping them altogether) so hurt performance when we used the initial selection from CommonCrawl. The fix this problem, we built a *diacritiser* to restore the correct diacritics to Romanian text, using clean Romanian (from Europarl) to generate a with/without diacritics parallel corpus. We then applied this diacritiser to our CommonCrawl selection. This made the synthetic data more effective, but also introduced noise in cases where the diacritiser gave bad results. The diacritiser was actually an NMT system itself, but perhaps a phrase-based system would have been more appropriate, as it would be more conservative. The noise in the en-ro synthetic data was also noted in *D2.3*.

After the initial success with the use of back-translated synthetic data, there are still many questions. In particular, we consider the following questions here:

- What is the best way to select data for back-translation?
- How should this data be combined with natural parallel data in training?
- How effective is it to increase the quantity of synthetic data and/or increase the proportion?

The experiments that follow will throw some light on these questions.

4.1.1 Selecting monolingual data

In the earlier experiments for the Y3 systems, monolingual data was selected either using the NHS 24 corpus or the Cochrane corpus. In the following experiments, we make a more detailed comparison of selection methods, and include a comparison with random selections.

We use a single language pair (en-cs) but train 4 independent runs in each condition, taking an ensemble of the best systems from each run. The baseline system uses the whole of the UFAL Medical Corpus,[12] which contains 49.6M sentences after applying standard Moses cleaning heuristics (rejecting lines longer than 80 tokens) and additionally removing any lines which contain no ASCII alphabetical characters. We learn a BPE model (Sennrich *et al.*, 2016b) on the parallel data using 89500 merge operations, and the updated heuristics from Sennrich *et al.* (2017) with a minimum occurrence count of 50. The translation

[11] There are other approaches, for example the semi-supervised models of He *et al.* (2016) and Cheng *et al.* (2016), which essentially train the forward and backwards systems simultaneously using parallel and monolingual data. We have implemented both of these approaches in Nematus, but have yet to see improvements over back-translation

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

models are trained using Nematus with the 4-layer deep recurrent transition network as used in Sennrich *et al.* (2017). We use the HimL 2015 tuning set as a validation set, validating every 10000 updates, and stopping training when validation cross-entropy fails to increase in 10 consecutive validation points.

After training the baseline system to convergence, we apply fine-tuning to the 4 different models separately, using different data sets. In each case we mix the adaptation data with the original parallel data, using the Nematus domain interpolation feature to achieve a 50-50 mix. We use the following selections of adaptation data:

EMEA The EMEA (new crawl) portion of the UFAL corpus.

COCHRANE Synthetic CommonCrawl selected using Cochrane.

NHS24 Synthetic CommonCrawl selected using NHS24.

BOTH Union of both synthetic selections above.

RANDOM Random selection from CommonCrawl.

In addition each of the above configurations (except the first) has a +EMEA version where it was combined with the EMEA corpus.

In Table 12 we show the BLEU scores on HimL 2015 test after fine-tuning with different selections of adaptation data.

Adaptation source	Adaptation Size	Cochrane	NHS24
baseline	0	33.2	25.5
EMEA	1.3M	34.3	26.5
COCHRANE	3.3M	34.3	26.8
COCHRANE+EMEA	4.5M	35.3	27.3
NHS24	3.9M	33.7	26.9
NHS24+EMEA	5.2M	**35.5**	27.6
BOTH+EMEA	8.4M	35.4	27.7
RANDOM	7.3M	33.4	26.8
RANDOM+EMEA	8.6M	34.6	**28.0**

Table 12: Comparison of selections of adaptation data, where we also show the size (in sentences) of the adaptation corpus (best results in bold). All adaptation is done by fine-tuning from baseline model. BLEU scores are for ensemble of 4 independent training runs on HimL 2015 test sets.

The first thing to note about the results in Table 12 is that it does not make much difference whether we use Cochrane or NHS24 to select data. Comparing the COCHRANE, COCHRANE+EMEA, NHS24 and NHS24+EMEA rows we can see that the performance on the NHS24 test set is slightly improved by using NHS24 data for selection. For Cochrane, the best performance of these four configurations is actually achieved using the NHS24+EMEA configuration, although only by +0.2 BLEU.

Fine-tuning with EMEA helps performance on both test sets, providing about a +1 BLEU gain. This is in contrast to earlier experiments, where we did not see a benefit in fine-tuning on EMEA, but in that case we were using just the de-duped EMEA adaptation set in the fine-tuning phase, as opposed to interpolating the general parallel corpus with the adaptation set, as we do now. In the earlier experiments, the adaptation set was too small, and the model quickly overfitted. The gains from EMEA consistently stack with gains from the synthetic corpora.

The performance with randomly selected data is curious. This was included as a way of checking whether the selection method from CommonCrawl is effective. For random selection, the only criterion is that we do not include sentences shorter than 10 tokens (and we exclude sentences longer than 50 from all data). For Cochrane there is little benefit in adding the randomly selected data from CommonCrawl (+0.2 or +0.3 BLEU). However for NHS24, adding the random selection to the adaptation set gives a gain of more than 1 BLEU in both cases (with or without EMEA).

One possibility is that BLEU is responding to better lengths of the translations. There does not seem to be a consistent pattern with the length penalties, but looking at Table 13 we can see that the mean lengths of the sentence in the synthetic corpora is much longer than that of the generic parallel data, and closer to the lengths in the test set.

Another possible explanation is that back-translated synthetic data only contains literal translations (in general) whereas naturally occurring parallel corpora may include non-literal translations, and also mis-aligned sentences. It has been noted elsewhere that removing semantically divergent sentence pairs from parallel corpora can help NMT (Carpuat *et al.*, 2017).

Corpus	Mean sent. length
General training	7.6
NHS24 test	14.2
Cochrane test	25.2
COCHRANE	17.3
NHS24	17.3
RANDOM	19.9
EMEA	15.3

Table 13: Mean sentence length (in tokens) of Czech part of each training and test corpus.

4.1.2 Mixing monolingual data

We now turn the the next question posed above about the adaptation data. How should it be mixed with the parallel data? So far we have used a fine-tuning approach, where a model is first trained to convergence on the parallel data, and then training is continued on a parallel/synthetic mix. The advantage of this approach is that we can (in theory) adapt the same generic system to multiple domains, however if we are training for a single domain then training time is increased. In (Sennrich *et al.*, 2017) we used this finetuning approach for some systems, but mostly we used a "mixed" approach, where the adaptation data is mixed with the natural parallel data from the start, and there is only a single phase of training.

Using the same experimental setup as above, we compare the following training regimes:

FINETUNE Train to convergence using the parallel data, then continue training using a 50-50 mix of adaptation and generic data. The mixing is done using the Nematus domain interpolation feature.

MIXED Mix the generic and adaptation data from the beginning of training, again using a 50-50 mix with Nematus's domain interpolation.

PREMIXED Instead of using domain interpolation, mix the two data sets in advance, oversampling the smaller one to create a 50-50 mix.

BATCH-NORM The batch normalisation method for mixing that was proposed in Wang *et al.* (2017), and found to be effective. It entails mixing the two data sets equally in each mini-batch, in contrast to the standard Nematus domain interpolation which mixes at the maxi-batch[13] level.

We compare these data mixing regimes using the BOTH+EMEA adaptation set from the previous experiment. In Table 14 we show the BLEU scores on HimL 2015 test, after training with the different regimes give above.

Mixing regime	Cochrane	NHS 24
FINETUNE	35.4	**27.7**
MIXED	35.2	27.3
PREMIXED	34.9	26.9
BATCH-NORM	**35.6**	26.2

Table 14: Comparison of data mixing regimes. BLEUscores are for ensemble of 4 independent training runs on HimL 2015 test sets.

The results in Table 14 are different for both HimL domains. For Cochrane, BATCH-NORM is the best technique, which is slightly better than FINETUNE, which in turn is slightly better than MIXED. For NHS 24, PREMIXED is again worse than the dynamic mixing options, but BATCH-NORM performs quite poorly. Based in these results, the best that can be said is that FINETUNE is the best on average, and MIXED is not far behind.

4.1.3 Scaling monolingual data

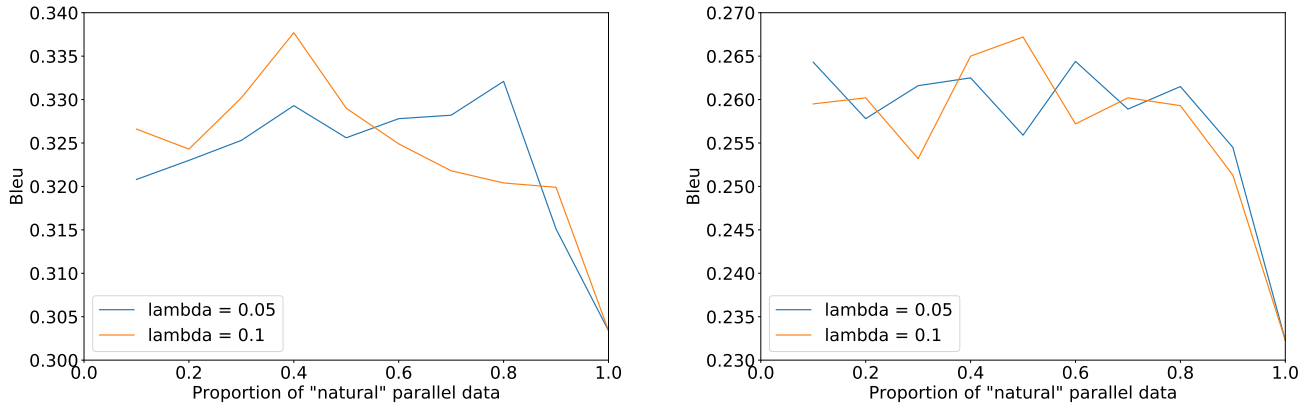
The experiments above used fixed amounts of synthetic data, mixed in a fixed proportion with the parallel data. We now look at the effect of varying both the amount of synthetic back-translated data, and the mixing proportion, using two different language pairs (en-cs and en-de).

[13]The default behaviour of Nematus is to load the data in maxi-batches, where each maxi-batch is the size of 20 mini-batches. Sentences are then sorted by length so that the maxi-batches are split into mini-batches of roughly equal length. This makes more efficient use of the GPU.

We prepare for these experiments by translating approximately 500 million sentences from the Czech and German CommonCrawls, into English. This represents the whole of the Czech CommonCrawl. Unlike before, we do not remove short sentences, but we do remove any sentence longer than 50 tokens, and any sentence containing no ASCII characters. Each target language sentence is scored using the Moore-Lewis scores derived from the translated Cochrane and NHS24 crawls, as described earlier. We fix a threshold λ (where $0 \leq \lambda \leq 1$) and, after ordering the sentences by ML score, select the λ proportion with the highest scores. This is done separately for both the Cochrane and NHS24 ML scores, and then we take the union (removing duplicates) of the two selected corpora to give us the adaptation corpora. For both language pairs, this selects about 75M synthetic parallel sentences with $\lambda = 0.1$.

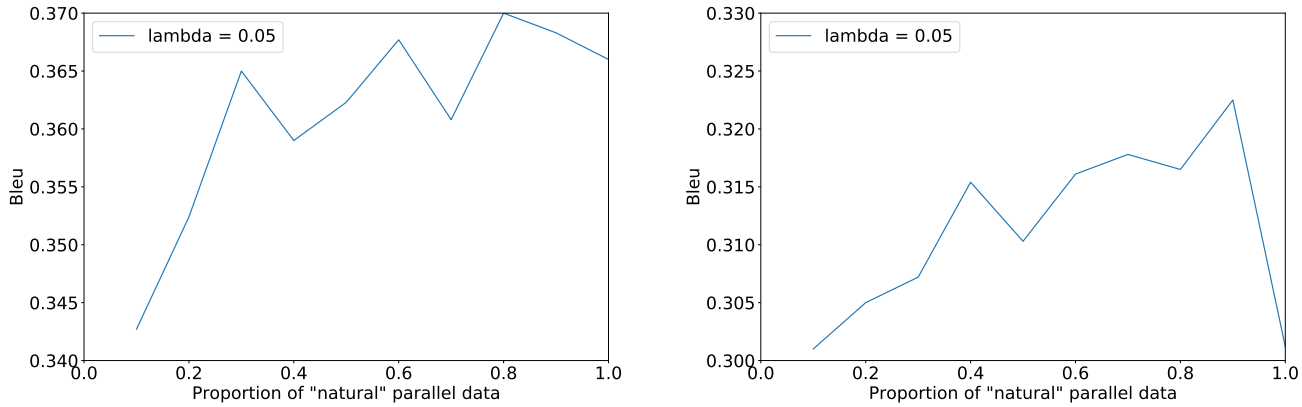
As well as the thresholding parameter λ, we also vary the mixing parameter μ. This parameter controls the proportion of naturally occurring parallel data in the training set, and is varied between 0.1 and 1.0 (where the latter means no synthetic data). For these experiments we use Marian (Junczys-Dowmunt *et al.*, 2016) as it is much faster to train, but since it does not support domain interpolation we use over-sampling to set the mixing proportions. That is to say, if one data set is smaller than it should be according to its mixing proportion, then it is repeated the appropriate number of times, with the final sentences being randomly sampled, to bring the proportions up to μ.

We train models with the shallow Nematus architecture, setting the Marian working memory (which determines the dynamic batch size) to 2500MB and otherwise using default parameters. As before, we use the UFAL medical corpus for training, and the HimL 2015 tuning sets for early-stopping. We show below the performance on the HimL 2015 test sets, varying μ and λ. The BLEU scores are for checkpoint ensembles (taking the final 4 checkpoints, at intervals of 30000).



(a) Cochrane **(b)** NHS 24

Figure 1: Bleu score versus mixing proportion (μ) on English→ Czech



(a) Cochrane **(b)** NHS 24

Figure 2: Bleu score versus mixing proportion (μ) on English→ German

From the graphs in Figure 1 we can see there is a clear effect from using *some* synthetic adaptation data. For NHS24, the mixing proportion (μ) matters less, there is just random fluctuation up to 0.8, and $\mu > 0.8$ gives poorer results. Raising the threshold to

0.1 gives the best performance, although the mixing proportion needs tuning. For Cochrane, again we see that $\mu > 0.8$ is bad and in fact for both values of λ we observe a definite peak, however at different values of μ. For the en-de pair, we see a different pattern, with little or no gain from synthetic data for Cochrane, but a significant (2 bleu point) gain for NHS24. In both cases a mixing proportion under 0.8 (i.e. adding too much synthetic data) causes a performance drop.

4.2 Using monolingual data without back-translation: An autoencoder for NMT

In this section we describe an alternative approach to using monolingual data in NMT, that does not require back-translation. The approach proved to be more suited to scenarios where there is very little naturally occurring parallel data, so we did not test directly on the HimL setups. The account given here is just a summary, with a full account available in (Currey *et al.*, 2017), included in Appendix B.

The idea is very simple. If you have monolingual data in the target language, then you can turn it into "parallel" data simply by using the same target data on the source side. This "copied monolongual data" is mixed in with the naturally occurring parallel data, and (optionally) synthetic data created by back-translation, and the NMT system is trained normally. In experiments on 6 language pairs (en↔tr, en↔ro and en↔de) we show improvements in all cases, except the pairs that include German. We suggest that this is because the technique works best when there is not much parallel data (the first two language pairs have under a million parallel sentences).

Analysis presented in the paper suggests that using copied monolingual data improves the translation accuracy on proper names and terms. These are often expected to be the same in source and target, or in some language pairs undergo minor changes due to the case system. However NMT generally does not have a pass-through mechanism, and out-of-vocabulary words can produce strange results, especially using subwords where OOVs are split down into known units. Using copied monolingual data means that the system attains better coverage of pass-through words at training time.

5 Conclusions

We have explored different ways in which monolingual data can be used to improve MT. The first idea was to use bilingual lexicon induction in order "fill in gaps" in the coverage of the parallel training corpus. We showed in Section 2 that there were indeed gaps in the coverage, despite the large training corpora, more pronounced in the case of Cochrane. In Section 3 we showed that we can make significant improvements over baseline approaches to bilingual lexicon induction for the medical domain.

The second way of using monolingual data is to create synthetic parallel data by back-translation, and use it in a neural MT system. We have shown that this provides consistent increases in bleu score, providing the data is selected appropriately, and gains can be increased by careful choice of the mixing parameters. Finally we showed a very simple way to use monolingual data for low-resource neural MT, just by copying it to the source to create "parallel" data.

References

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. "Enriching word vectors with subword information." *CoRR*, abs/1607.04606.

Buck, Christian, Kenneth Heafield, and Bas van Ooyen. 2014. "N-gram counts and language models from the common crawl." *Proceedings of the Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Carpuat, Marine, Yogarshi Vyas, and Xing Niu. 2017. "Detecting cross-lingual semantic divergence for neural machine translation." *Proceedings of the First Workshop on Neural Machine Translation*, 69–79. Vancouver.

Cheng, Yong, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. "Semi-supervised learning for neural machine translation." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1965–1974. Berlin, Germany.

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

Duong, Long, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. 2016. "Learning crosslingual word embeddings without bilingual corpora." *Proc. EMNLP*.

Durrani, Nadir, Barry Haddow, Philipp Koehn, and Kenneth Heafield. 2014. "Edinburgh's phrase-based machine translation systems for wmt-14." *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 97–104. Baltimore, Maryland, USA.

Dyer, Chris, Victor Chahuneau, and Noah A Smith. 2013. "A Simple, Fast, and Effective Reparameterization of IBM Model 2." *Proceedings of NAACL*.

Faruqui, Manaal and Chris Dyer. 2014. "Improving vector space word representations using multilingual correlation." *Proc. EACL*.

Gouws, Stephan and Anders Søgaard. 2015. "Simple task-specific bilingual word embeddings." *Proc. NAACL*.

He, De., Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. "Dual Learning for Machine Translation." *Proceedings of NIPS*. Also in NIPS 2016.

Heyman, Geert, Ivan Vulić, and Marie-Francine Moens. 2017. "Bilingual lexicon induction by learning to combine word-level and character-level representations." *Proceedings of EACL 2017*, 1084–1094.

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. "Is neural machine translation ready for deployment? a case study on 30 translation directions." *Proceedings of the 9th International Workshop on Spoken Language Translation (IWSLT)*. Seattle, WA.

Lazaridou, Angeliki, Georgiana Dinu, and Marco Baroni. 2015. "Hubness and pollution: Delving into cross-space mapping for zero-shot learning." *Proc. ACL*.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. "Efficient estimation of word representations in vector space." *Proceedings of Workshop at ICLR*.

Mikolov, Tomas, Quoc V Le, and Ilya Sutskever. 2013b. "Exploiting similarities among languages for machine translation." *CoRR*, abs/1309.4168.

Sennrich, Rico, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. "The university of edinburgh's neural mt systems for wmt17." *Proceedings of the Second Conference on Machine Translation*, 389–399. Copenhagen, Denmark.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. "Improving neural machine translation models with monolingual data." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. Berlin, Germany.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. "Neural machine translation of rare words with subword units." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany.

Vulic, Ivan and Anna Korhonen. 2016. "On the Role of Seed Lexicons in Learning Bilingual Word Embeddings." *Proc. ACL*, 247–257.

Wang, Rui, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. "Instance weighting for neural machine translation domain adaptation." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1483–1489. Copenhagen, Denmark.

Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin. 2015. "Normalized word embedding and orthogonal transform for bilingual word translation." *Proc. NAACL*.

A Coverage Tables

This appendix shows importance-coverage data for incorrectly translated terms for each language-domain pair in HimL. The tables are sorted by increasing coverage, and cut off so that each table fits onto a single page. References translations of terms are extracted by projecting across automatic alignments of the test sets. Full definitions of coverage, importance and count are given in Section 2, but in brief they are defined as:

coverage The maximum coverage over all possible segmentations of a term, where the coverage of a segmentation is the smallest count of each segment in the training corpus.

importance The difference between the log-probability under an in-domain language model versus an out-of-domain language model.

count The number of times the term is contained in the training set.

Source	Importance	Coverage	Hypothesis	Reference	Count
Okwundu CI	1.881	0	Okwundu CI	CI Okwundu	1
assay EXTEM clot amplitude	3.217	0	assay EXTEM sraženina amplitudy	amplituda sraženiny testu EXTEM	1
ECSW and internal fixation	1.499	0	ECSW a vnitřní fixaci	ECSW a vnitřní fixací	1
Tuinebreijer WE	1.731	0	Tuinebreijer WE	Tuinebreijer MY	1
rotational thromboelastometry	2.590	0	rotační thromboelastometry	rotační tromboelastometrie	5
LIPUS and control	1.323	0	Lipus a kontroly	metodou Lipus a kontrolní	1
subfertile women	1.961	0	subfertile žen	subfertilních žen	1
CA10 and CA15 measurements	2.741	0	CA10 a ca 15 měření	CA10 a CA15	1
CI -22.71	1.311	0	CI -22.71	CI -22,71	1
handsearching	1.121	0	handsearching	ručně	2
Kisely SR	1.782	0	Kisely SR	SR Kisely	1
Bossuyt PMM	2.061	0	BOSSUYT PMM	Bossuyt PMM	1
Laopaiboon M	1.427	0	Laopaiboon M	M Laopaiboon	1
QUADAS-2 tool	2.295	1	nástroje nemůžeš-2	nástroje QUADAS-2	1
CI -1.14	1.391	1	CI -1.14	CI -1,14	1
CA5 measurements and 1	1.678	1	měření CA5 a 1	měření CA5 a 1 měření	1
uninterpretable Rotem study results	2.408	1	stala naprosto nesrozumitelnou Rotem studijní výsledky	žádné zahrnutých studií výsledky ROTEM	1
Zhelev Z	1.650	1	Želev Z	Zhelev Z	1
Rotem test	1.165	2	Rotem test	testu ROTEM testu	1
TEG or Rotem	2.226	2	TEG nebo Rotem	TEG nebo ROTEM	2
TEG and Rotem assessments	2.565	2	TEG a Rotem hodnocení	hodnocení TEG a ROTEM	1
centres TEG and Rotem	2.610	2	střediscích TEG a Rotem	centrech TEG a ROTEM	1
TEG and Rotem)	2.004	2	TEG a Rotem)	TEG a ROTEM)	1
Rotem assessment	1.347	2	Rotem hodnocení	hodnocení ROTEM	1
Rotem and TEG	1.782	2	Rotem a TEG	ROTEM i TEG	1
Rotem CAs	1.891	2	Rotem CA	ROTEM CA	1
TEG and Rotem	2.296	2	Teg a Rotem	Teg a ROTEM	1
			TEG a Rotem	TEG a ROTEM	2
coagulopathic trace	1.942	3	coagulopathic stopy	koagulopatického sledování	1
TEG assessment	1.591	4	TEG hodnocení	hodnocení TEG	1
CI -0.28	1.554	6	interval spolehlivosti -0,28	CI -0,28	1
dichotomous outcomes	2.660	8	dichotomické výsledky	dichotomní výsledky	2
new RCTs	1.240	8	nové studie	nových RCT	1
Peto odds ratio	2.210	8	Peto odds ratio	poměr šancí Petovou	1
bias RCTs	1.719	8	zkreslení RCTs	RCT zkreslení	1
single RCT	2.054	8	jedné koridorové	jediné RCT	1
dichotomous outcome	2.332	8	dichotomické výsledku	dichotomního výsledku	1
CI -1.23	1.555	8	CI -1.23	CI -1,23	1
further RCT	1.473	8	další ošetření	další RCT	1
Specialised Register	2.864	13	Specializované registru	registr Cochranovy	1
CI 2.08	1.284	21	interval spolehlivosti 2,08	CI 2,08	1

Table 15: Incorrectly translated domain-specific terms for cochrane English-Czech

Source	Importance	Coverage	Hypothesis	Reference	Count
assay EXTEM clot amplitude	3.217	0	Assay EXTEM Gerinnsel Amplitude	Analyse extem Gerinnselamplitude CA	1
ECSW and internal fixation	1.499	0	ECSW und Osteosynthesen	ECSW und interner Fixation	1
rotational thromboelastometry	2.590	0	rotatorische thromboelastometry	rotatorischer Thromboelastometrie	1
			thromboelastometry	eine rotatorische Thromboelastometrie	1
			rotatorische thromboelastometry	rotatorische Thromboelastometrie	1
			thromboelastometry	rotatorische Thromboelastometrie	2
LIPUS and control	1.323	0	LIPUS und Kontrolle	LIPUs und Kontrolle	1
CI -22.71	1.311	0	KI -22.71	KI -22,71	1
Laopaiboon M	1.427	0	Lumbiganon M	Laopaiboon M	1
EBSCO CINAHL	2.575	0	Datenbankensysteme EBSCO CINAHL	EBSCO CINAHL	1
subfertile women	1.961	1	Frauen Fertilitätsstörungen	subfertilen Frauen	1
QUADAS-2 tool	2.295	1	QUADAS-2 Werkzeug	Instruments QUADAS-2	1
CI -1.14	1.391	1	KI -1.14	KI -1,14	1
uninterpretable Rotem study results	2.408	1	uninterpretable Rotem Studienergebnisse	undeutbaren ROTEM Studienergebnisse	1
Rotem and TEG	1.782	5	Rotem und TEG	ROTEM und TEG	1
TEG and Rotem)	2.004	5	TEG und Rotem)	TEG und ROTEM)	1
TEG and Rotem assessments	2.565	5	TEG und Rotem Bewertungen	TEG und ROTEM Bewertungen	1
centres TEG and Rotem	2.610	5	Zentren TEG und Rotem	Zentren TEG und ROTEM	1
TEG and Rotem	2.296	5	TEG und Rotem	TEG und ROTEM	3
Rotem CAs	1.891	5	Rotem CAs	ROTEM CAs	1
Rotem assessment	1.347	5	Rotem Bewertung	ROTEM Bewertung	1
Rotem test	1.165	5	Rotem-Test	ROTEM-Tests	1
TEG or Rotem	2.226	5	TEG oder Rotem	TEG oder ROTEM	1
			TEG oder Rotem	TEG- oder ROTEM	1
coagulopathic trace	1.942	8	koagulopathischen Spur	koagulopathischen Spuren	1
bias RCTs	1.719	13	Bias RCTs	RCTs Verzerrungen	1
new RCTs	1.240	13	neue randomisierte kontrollierte Studien	neue RCTs	1
Peto odds ratio	2.210	17	Peto Odds Ratio	Peto Odds-Ratio	1
dichotomous outcome	2.332	50	dichotome Ergebnis	dichotomen Endpunkt	1
dichotomous outcomes	2.660	50	dichotome Endpunkte	dichotome Ergebnisse	2
repping S	1.148	50	vertrete ein S	Repping S	1
soldiers or midshipmen	1.914	80	Soldaten oder Kadett waren	Soldaten oder Seekadetten	1
Diagnostic Test Accuracy Reviews	3.261	98	diagnostische Test Accuracy Reviews	Diagnostic Test Accuracy Reviews	1
TEG assessment	1.591	127	Bewertung TEG	TEG Bewertung	1
PTR / INR reading	2.669	132	PTR / INR Lesen	PTR / INR Messwert	1
PTR / INR test	2.562	132	PTR / INR Test	PTR / INR-Wert-Tests	1
median Injury Severity Scores	1.694	137	Median Injury Severity Score	mediane Injury Severity Scores	1
overestimating benefits	1.747	163	überschätzen Nutzen	Nutzen	1
RR 0.49	1.967	211	RR = 0,49	RR 0,49	1
Senior Peer Researcher	2.797	220	Senior Peer Researcher	Senior Peer-Forscher	1
CI 0.79	2.022	238	KI 0.79	KI 0,79	1

Table 16: Incorrectly translated domain-specific terms for cochrane English-German

Source	Importance	Coverage	Hypothesis	Reference	Count
assay EXTEM clot amplitude	3.217	0	assay EXTEM skrzepu amplitudy	amplituda skrzepu analizy EXTEM	1
Glujovsky D	1.082	0	Glujovsky D	D. Glujovsky	1
LIPUS and control	1.323	0	LIPUS i kontroli	LIPUS a kontrolną	1
Ukoumunne O	1.166	0	Ukoumunne O	O. Ukoumunne	1
rotational thromboelastometry	2.590	0	thromboelastometry obrotowa	tromboelastometria rotacyjna	1
			rotacyjnej thromboelastometry	tromboelastometria rotacyjna	2
			thromboelastometry obrotowa	Rotem	2
ECSW and internal fixation	1.499	0	ECSW i wewnętrznie	pozaustrojowej terapii uderzeniowej	1
Ciapponi A	1.405	0	Ciapponi A	A. Ciapponi	1
Tuinebreijer WE	1.731	0	Tuinebreijer WE	W.E. Tuinebreijer	1
CI -22.71	1.311	0	CI -22.71	przedział ufności CI -22,71	1
Okwundu CI	1.881	0	Okwundu CI	C.I. Okwundu	1
handsearching	1.121	0	handsearching	ręcznie	1
			handsearching	Medline	1
Kisely SR	1.782	0	Kisely SR	S.R. Kisely	1
Laopaiboon M	1.427	0	Laopaiboon M	M. Laopaiboon	1
Thanaviratananich S	1.701	0	Thanaviratananich S	s. Thanaviratananich	1
RIESTRA B	1.332	1	Riestra B	B. Riestra	1
subfertile women	1.961	1	subfertile kobiet	kobiet niepłodnością	1
CI -1.14	1.391	1	CI -1.14	przedział ufności CI -1,14	1
uninterpretable Rotem study results	2.408	1	uninterpretable Rotem wyniki badań	żadnym badań wynikach Rotem	1
Kerse N	1.270	1	Kerse N	N. Kerse	1
QUADAS-2 tool	2.295	2	QUADAS-2 narzędzie	narzędzia QUADAS-2	1
CI -2.14	1.671	2	CI -2.14	przedział ufności CI -2,14	1
Bossuyt PMM	2.061	2	BOSSUYT PMM	P.M.M. Bossuyt	1
Ovid EMBASE	2.595	3	Ovid Embase	UNALIGNED	1
CA10 and CA15 measurements	2.741	5	CA10 i antygen pomiarów	CA10 i CA15	1
Rotem CAs	1.891	8	Rotem CA	amplitud skrzepu Rotem	1
Rotem assessment	1.347	8	Rotem oceny	oceną Rotem	1
TEG and Rotem assessments	2.565	8	TEG i Rotem oceny	TEG i Rotem koagulopatii	1
centres TEG and Rotem	2.610	8	ośrodków TEG i Rotem	ośrodkach TEG i Rotem	1
Rotem test	1.165	8	Rotem badania	test Rotem	1
CA5 measurements and 1	1.678	10	Ca5 pomiarów i 1	pomiaru Ca5 i	1
CI -1.23	1.555	11	CI -1.23	przedział ufności CI -1,23	1
bias RCTs	1.719	13	stronniczości RCTs	randomizowanych błędu systematycznego	1
RCTs	2.295	13	RCTs	randomizowanych badaniach klinicznych (RCT)	1
			RCTs	badań	1
			RCTs	randomizowanych badań klinicznych	1
			RCTs	badania	1
			RCTs	randomizowanych badaniach klinicznych	1
			RCTs	uwzględniliśmy randomizowane kontrolowane badania kliniczne	1

Table 17: Incorrectly translated domain-specific terms for cochrane English-Polish

Source	Importance	Coverage	Hypothesis	Reference	Count
Okwundu CI	1.881	0	IÎ Okwundu	Okwundu C.I.	1
assay EXTEM clot amplitude	3.217	0	testul EXTEM cheag amplitudine	testul extem pentru amplitudinea cheagului	1
LIPUS and control	1.323	0	LIPUS și control	LIPUS și	1
ECSW and internal fixation	1.499	0	ECSW și fixare internă	internă ECSW și fixarea internă	1
rotational thromboelastometry	2.590	0	thromboelastometry de rotație	trombelastometrie rotațională	1
			de rotație thromboelastometry	trombelastometria rotațională	4
CI -22.71	1.311	0	IÎ -22.71	IÎ -22.71	1
handsearching	1.121	0	handsearching	UNALIGNED	2
Kisely SR	1.782	0	Kisely SR	Kisely S.R.	1
Thanaviratananich S	1.701	0	Thanaviratananich S	Thanaviratananich S.	1
CI -1.14	1.391	1	IÎ -1.14	IÎ -1.14 0.20	1
subfertile women	1.961	1	subfertile femei	femeile recurg	1
uninterpretable Rotem study results	2.408	1	uninterpretable îmbătrânește rezultatele studiului	au rezultatele studiului ROTEM	1
Kerse N	1.270	1	Kerse N	Kerse N.	1
CA10 and CA15 measurements	2.741	2	măsurarea CA10 și CA15	și măsurarea AC10 și AC15	1
QUADAS-2 tool	2.295	2	instrument QUADAS 2	instrumentul QUADAS-2	1
Ovid EMBASE	2.595	3	Ovidiu EMBASE	Ovid EMBASE	1
TEG and Rotem	2.296	6	TEG și îmbătrânește	TEG și ROTEM	3
Rotem CAs	1.891	6	îmbătrânește ’	de ROTEM	1
Rotem test	1.165	6	testului îmbătrânește	testului ROTEM testul	1
TEG and Rotem assessments	2.565	6	TEG și evaluările îmbătrânește	TEG și ROTEM în	1
centres TEG and Rotem	2.610	6	centre TEG și îmbătrânește	centre TEG și ROTEM	1
Rotem assessment	1.347	6	evaluarea îmbătrânește	ROTEM	1
TEG or Rotem	2.226	6	TEG sau îmbătrânește	TEG sau ROTEM	1
			TEG sau îmbătrânește	TEG ROTEM	1
TEG and Rotem)	2.004	6	TEG și îmbătrânește)	TEG și ROTEM)	1
Rotem and TEG	1.782	6	îmbătrânește și TEG	ROTEM și TEG	1
CA5 measurements and 1	1.678	10	măsurarea CA5 și 1	măsurarea AC5	1
CI -1.23	1.555	11	IÎ -1.23	IÎ -1.23	1
RCTs	2.295	13	RCTs	studii clinice randomizate controlate	1
			RCTs	CHM	1
			RCTs	studii randomizate controlate	1
			RCTs	studiile clinice randomizate controlate	1
			RCTs	SCR	3
			RCTs	SCR-uri	6
			RCTs	SCR uri	14
bias RCTs	1.719	13	părtinire RCTs	SCR părtinire	1
new RCTs	1.240	13	nou RCTs	SCR uri noi	1
CI -0.28	1.554	13	CI -0.28	IÎ -0.28 0.09	1
coagulopathic trace	1.942	14	urme coagulare	coagulopatiei	1
Peto odds ratio	2.210	24	Peto relativ	rație Peto	1

Table 18: Incorrectly translated domain-specific terms for cochrane English-Romanian

Source	Importance	Coverage	Hypothesis	Reference	Count
eatwell Plate	1.582	0	EATWELL Plate	potravin	1
bifocal or varifocal lenses	1.834	0	bifokály nebo varifocal čočky	bifokálními nebo varifokálními čočkami	1
RNIB Scotland	1.548	0	KNIN Skotsko	RNIB Skotsko	2
eatwell plate	1.390	0	talíř EATWELL	talíři správnou skladbou	1
bifocals or varifocals	1.980	0	bifokály nebo varifocals	bifokálními nebo varifokálními	1
Dosette ’	1.007	0	Dosette "	dávkovače léků neboli	1
130 / 80mmHg	1.794	0	130 / 80mmHg	130 / 80 mmHg	1
120 / 80mmHg	1.849	0	120 / 80mmHg	120 / 80 mmHg	3
front knee strengthener	2.400	2	přední roborans koleno	posilovač přední strany kolen	1
kerbs and steps	1.244	10	obrubníky a kroky	obrubníky	1
kerbs or steps	1.613	10	obrubníky nebo kroků	obrubníků či schodů	1
frailer older people	1.496	21	subtilnějších starších lidí	starších lidí křehkého kteří	1
electronic Medicines Compendium	2.631	27	elektronického Compendium pro léčivé přípravky	elektronického kompendia	1
tinned pilchards	1.986	28	konzervované sardinky	sardinky konzervě	1
sardines and pilchards	1.759	28	sardinky a sardinky	i sardinky	1
Friday 8.45 AM	1.803	42	pátek 20.45 AM	pátek 8 : 45	1
physiotherapist or occupational therapist	1.701	46	fyzioterapeuta nebo terapeute	fyzioterapeuta nebo rehabilitačního pracovníka	1
			fyzioterapeuta či terapeute	fyzioterapeutem nebo rehabilitačním ,	1
physiotherapist or exercise specialist	2.060	46	fyzioterapeut nebo cvičení specialistou	cvičení fyzioterapeutem nebo odborníkem	1
lightheadedness and tiredness	1.969	49	závratě a únava	točení hlavy a únava	1
GP or optometrist	1.235	78	lékařem nebo oftalmologem	praktickému lékaři nebo optometrikovi	1
near and farsighted people	1.810	79	poblíž a prozíravých lidí	tak dalekozraké lidi	1
5.30 PM	1.291	90	17 : 30 hod.	17 : 30	1
helpline	1.124	94	linky pomoci	linku pomoci	2
Consultant cardiologist	2.087	109	Consultant Kardiolog	kardiologa	1
uneven pavements	2.013	114	nerovným chodníky	nerovnými chodníky	1
small matchbox size piece	2.755	122	malé velikosti krabičky kus	kousek velikosti malé krabičky zápalek	1
			malé velikosti krabičky zápalek kus	kousek velikosti malé krabičky od zápalek	1
gym memberships	1.201	134	fittek	slevu členství	1
cabbage and okra	1.712	136	zelí a okru	zelí a okře	1
NHS eye tests	1.576	148	NHS oční testy	oční vyšetření zdravotní péče	1
NHS Scotland	1.736	148	NHS Skotsko	systému Skotska	1
NHS	3.330	148	zdravotnictví	rámci systému státní zdravotní péče	1
			NHS	SzS	27
NHS Low Income Scheme	2.431	148	NHS nízkým příjmem Scheme	nízkým systému státní zdravotní péče	1
controllable lighting levels	1.382	173	kontrolovatelná hladiny osvětlení	ovladatelné osvětlení	1
rubber stoppers or wheels	1.666	190	pryžové zátky nebo kola	gumových nástavců koleček	1
balloon (angioplasty)	1.511	196	balónem (angioplastika)	balónku (angioplastika)	1
equipment and adaptations	1.342	225	vybavení a úpravy	vybavení a úpravách	1
			zařízení a úpravy	s opravami , vylepšeními a úpravami	1
Friday 9.00 AM	1.366	227	pátek 9.00 AM	pátek 9.00	1

Table 19: Incorrectly translated domain-specific terms for nhs24 English-Czech

Source	Importance	Coverage	Hypothesis	Reference	Count
bifocals or varifocals	1.980	0	Bifokal- oder varifocals	Bifokal- oder Varifokalgläser	1
Dosette ’	1.007	0	Dosette "	Pillen	1
eatwell Plate	1.582	1	eatwell Plate	Eatwell Platte	1
eatwell plate	1.390	1	eatwell Platte	Eatwell Platte	1
RNIB Scotland	1.548	5	RNIB Schottland	RNIB Scotland	2
frailer older people	1.496	23	frailer ältere Menschen	hinfällige ältere Bewohnern	1
kerbs or steps	1.613	34	Kerben oder Schritten	Bordsteine oder Treppen	1
bifocal or varifocal lenses	1.834	34	bifokale oder varifokalen Linsen	Brille Bifokal- oder Varifokalgläsern	1
kerbs and steps	1.244	34	Bordsteinkanten und Schritte	begehen Bordsteinkanten und Stufen	1
tinned pilchards	1.986	37	verzinnt Sardinen-	Sardinen	1
sardines and pilchards	1.759	37	Sardinen und Sardinen-	Sardinen und Pilchards	1
front knee strengthener	2.400	51	vorderen Knie Filmverstärker	Stärkung des vorderen Knies	1
stamina and suppleness	1.502	93	Ausdauer und Geschmeidigkeit	Ausdauer und Gelenkigkeit	1
NHS Choices	2.289	95	der NHS Choices	NHS Choices	1
GP or optometrist	1.235	112	Arzt oder Optiker	Hausarzt oder Optiker	1
Friday 8.45 AM	1.803	121	Freitag 20.45 Uhr	Freitag 8.45 Uhr	1
physiotherapist or exercise specialist	2.060	132	Physiotherapeuten oder körperliche Bewegung Spezialisten	Physiotherapeuten oder Trainingsexperten Sie	1
physiotherapist or occupational therapist	1.701	132	Physiotherapeut oder Beschäftigungstherapeut	Physiotherapeuten oder einen Beschäftigungstherapeuten zu besuchen	1
			Physiotherapeuten oder Beschäftigungstherapeut	Physiotherapeuten oder Beschäftigungstherapeuten	1
small matchbox size piece	2.755	165	kleine Matchbox Größe Stück	ein Stück Größe Streichholzschachtel	1
			Streichholzschachtel Größe kleines Stück	kleines Stück (groß Streichholzschachtel	1
near and farsighted people	1.810	193	in der Nähe und weitsichtig Menschen	und Weitsichtigkeit	1
monounsaturated fats	2.039	288	monoungesättigte Fette	einfach ungesättigte Fettsäuren	1
gym memberships	1.201	311	Fitnessmitgliedschaften	Fitnessstudio Mitglieder	1
Dr Peter Henriksen	1.897	318	Dr Peter Henriksen	Dr. Peter Henriksen	1
uneven pavements	2.013	353	unebene Bürgersteige	ungleichmäßige Pflaster	1
Consultant cardiologist	2.087	374	Consultant Kardiologen	Facharzt Kardiologie	1
Friday 9am	1.619	377	freitags 9 bis	Freitag von 9 bis	1
			Freitag 9 bis	Freitag von 9 bis	1
NHS eye tests	1.576	432	NHS Auge Tests	NHS Sehtests für	1
NHS Low Income Scheme	2.431	432	NHS einkommensschwache Scheme	NHS Low Income Scheme	1
NHS Scotland	1.736	432	NHS Schottland	NHS Scotland	1
electronic Medicines Compendium	2.631	463	elektronischen Medicines Kompendium	elektronischen Arzneimittel Kompendiums	1
light fingertip touch	2.318	475	leichten Fingerspitze berühren	leicht mit den Fingerspitzen	1
normal-strength lager	1.771	550	normal festen Lagerbier	Bier	1
more purposeful march	1.899	696	gezielteren Marsch	zweckmäßiger marschieren	1
slightly unsteady	1.716	698	etwas unsicher	wenig unsicher	1
regular , nutritious meals	1.579	716	regelmäßig , nahrhaftes Essen	regelmäßiges , nahrhaftes Essen	1
active housework	1.454	746	aktive Hausarbeit	aktive Haushaltsarbeiten	1
5.30 PM	1.291	790	17.30 Uhr	5.30 Uhr	1
soya beans and tofu	1.381	928	Sojabohnen und Tofu	Soja und Tofu	1

Table 20: Incorrectly translated domain-specific terms for nhs24 English-German

Source	Importance	Coverage	Hypothesis	Reference	Count
bifocals or varifocals	1.980	0	dwuogniskowce lub varifocals	dwu- lub zmiennoogniskowymi okularami albo rozpatrujesz zakup	1
130 / 80mmHg	1.794	0	130 / 80mmHg	130 / 80 mmHg	1
Dosette ’	1.007	0	Dosette "	" Dozownik "	1
eatwell plate	1.390	3	dobrzejeść płytki	piramida zdrowego odżywiania	1
eatwell Plate	1.582	3	dobrzejeść Plate	odżywiania Talerz	1
RNIB Scotland	1.548	7	RNIB Szkocji	RNIB Scotland	2
frailer older people	1.496	27	wątli starszych osób	starszych osób kruchej	1
bifocal or varifocal lenses	1.834	34	bifocal lub varifocal soczewki	szkłami dwu- albo zmiennoogniskowymi	1
kerbs and steps	1.244	43	krawężniki i czynności	stopniach z	1
kerbs or steps	1.613	43	krawężniki lub kroków	schodzeniu z progu	1
front knee strengthener	2.400	53	kolano strengthener przed	wzmacnianie stawu kolanowego	1
lightheadedness and tiredness	1.969	83	zawroty głowy i uczucie zmęczenia	zawroty głowy i zmęczenie	1
stamina and suppleness	1.502	100	wytrzymałość i suppleness	i równowagi wytrzymałości	1
Heart Helpline	1.813	110	serce Linia	infolinię Heart helpline	1
			serce Linia	Heart helpline numer	1
			serce Linia	Heart helpline	3
sardines and pilchards	1.759	110	sardynki i sardele	sardynki sardele	1
NHS Choices	2.289	174	NHS wybory "	NHS Choices	1
Friday 8.45 AM	1.803	206	piątku 8.45 AM	piątku 8.45	1
physiotherapist or exercise specialist	2.060	229	fizjoterapeuty lub wykonywania specjalistyczne	fizjoterapeutą lub specjalistą od ćwiczeń	1
physiotherapist or occupational therapist	1.701	229	fizjoterapeuty lub terapeuta zajęciowy	u fizjoterapeuty lub terapeuty zajęciowego	1
			fizjoterapeuty lub terapeuta zajęciowy	fizjoterapeutą lub terapeutą zajęciowym	1
caffeinated drinks	2.237	261	kofeiną napoje	napojów kofeinowych	1
			kofeiną picie	napoje kofeiną	1
GP or optometrist	1.235	311	lekarzem lub optyka	lekarzem rodzinnym czy okulistą	1
monounsaturated fats	2.039	332	tłuszcze nienasycone	nienasycone tłuszcze	1
near and farsighted people	1.810	347	w pobliżu i dalekowidzem ludzi	krótkowidzów i dalekowidzów	1
helpline	1.124	398	zaufania	infolinię	1
			zaufania	infolinii	1
Dr Peter Henriksen	1.897	413	dr Peter Henriksen	doktora Petera Henriksena	1
small matchbox size piece	2.755	428	małe zapałek rozmiar sera	kawałek rozmiaru pudełka zapałek	1
			małe zapałek rozmiar sera	porcja kawałku wielkości pudełka zapałek	1
gym memberships	1.201	449	siłownię członkostwa	członkostwo siłowni	1
uneven pavements	2.013	482	nierówne chodników	nierównych chodnikach	1
Consultant cardiologist	2.087	535	Consultant Cardiologist	kardiologa	1
breathlessness	1.041	595	duszność	brak tchu	1
light fingertip touch	2.318	611	świetle dotykać palcami	delikatnie dotykać palcami	1
cabbage and okra	1.712	636	kapusty i piżmian	kapuście okrze	1
Friday 9am	1.619	683	piątku 9	piątku 9	2
NHS eye tests	1.576	711	NHS oka badania	badanie wzroku NHS	1
NHS Low Income Scheme	2.431	711	NHS o niskich dochodach "	NHS Low Income Scheme	1

Table 21: Incorrectly translated domain-specific terms for nhs24 English-Polish

Source	Importance	Coverage	Hypothesis	Reference	Count
bifocals or varifocals	1.980	0	lentilele contact sau varifocals	lentilele bifocale sau varifocale	1
Dosette ’	1.007	0	Dosette "	" Dosette "	1
eatwell plate	1.390	2	platou eatwell	diagrama mâncatului sănătos	1
eatwell Plate	1.582	2	eatwell Plate	Eatwell Plate	1
frailer older people	1.496	23	frailer mai în vârstă persoane	persoane vârstnice mai firave care	1
bifocal or varifocal lenses	1.834	34	lentile bifocal sau varifocal	lentile bifocale sau varifocale	1
kerbs or steps	1.613	39	utilizati borduri sau etape	borduri sau trepte	1
kerbs and steps	1.244	39	utilizati borduri și măsurile	treptelor primejdii și	1
front knee strengthener	2.400	52	genunchi întărește față	întărirea genunchiului	1
sardines and pilchards	1.759	66	sardine și sardele	sardinele și sardele	1
tinned pilchards	1.986	66	conserve de sardine	conservă de sardele	1
lightheadedness and tiredness	1.969	78	amețeli și oboseală	moleșeală și oboseală	1
Heart Helpline	1.813	82	" Linia	Heart helpline	1
			Heart Linia	Heart helpline	4
stamina and suppleness	1.502	100	rezistența și agilitate	vitalitatea și suplețea	1
NHS Choices	2.289	130	NHS Alegerile	SNS Alegeri	1
Friday 8.45 AM	1.803	160	vineri la 8 : 45 AM	vineri 8.45	1
physiotherapist or exercise specialist	2.060	178	fizioterapeut sau exercită specialist	fizioterapeut sau specialist	1
GP or optometrist	1.235	227	medicul sau un oftalmolog	medicul sau un optometrist	1
near and farsighted people	1.810	275	în apropiere de și prezbit	persoanele de aproape și departe	1
swimming and hydrotherapy	1.558	284	înot și hidroterapie	înotul și hidroterapia	1
monounsaturated fats	2.039	301	mononesaturate grăsimi	grăsimi mononesaturate	1
small matchbox size piece	2.755	310	cutie de chibrituri dimensiune mică bucată	bucată mărimea unei cutii mici chibrituri	1
			de chibrituri dimensiune mică bucată	bucată de mărimea unei cutii mici chibrituri	1
Dr Peter Henriksen	1.897	383	dr Peter Henriksen	dr . Peter Henriksen	1
gym memberships	1.201	387	legitimațiilor	abonamentele de sală celor	1
uneven pavements	2.013	410	inegal pardoseli	trotuarele neregulate	1
Consultant cardiologist	2.087	473	Consultant Cardiolog	cardiolog consultant	1
electronic Medicines Compendium	2.631	514	Medicamente Compendium electronice	Compendiul Medicamente electronic	1
breathlessness	1.041	530	senzația aer	senzația	1
light fingertip touch	2.318	549	lumină vârful degetului atinge	atingi ușor vârful degetului	1
Friday 9am	1.619	558	vineri 9 dimineața	vineri la 9 la	2
NHS Scotland	1.736	576	NHS Scoția	NHS Scotland	1
NHS eye tests	1.576	576	testele ochi NHS	testele oftalmologice SNS	1
NHS Low Income Scheme	2.431	576	NHS cu Scheme	pentru Venituri Reduse SNS (	1
NHS Fife	2.095	576	NHS Fife	SNS Fife	1
lightheaded and dizzy	1.495	620	amețeală și amețeală	moleșit sau amețit	1
more purposeful march	1.899	834	marș mai	mers mai	1
normal-strength lager	1.771	855	bere normale de putere	bere obișnuită shot tărie	1
lager or cider	2.032	855	bere sau suc de mere	lager sau cidru	1
slightly unsteady	1.716	898	ușoară instabil	o ușoară	1

Table 22: Incorrectly translated domain-specific terms for nhs24 English-Romanian

B Paper: Copied Monolingual Data Improves Low-Resource Neural Machine Translation

Copied Monolingual Data Improves Low-Resource Neural Machine Translation

Anna Currey, Antonio Valerio Miceli Barone, and **Kenneth Heafield**
School of Informatics, University of Edinburgh
a.currey@sms.ed.ac.uk
{amiceli, kheafiel}@inf.ed.ac.uk

Abstract

We train a neural machine translation (NMT) system to both translate source-language text and copy target-language text, thereby exploiting monolingual corpora in the target language. Specifically, we create a bitext from the monolingual text in the target language so that each source sentence is identical to the target sentence. This copied data is then mixed with the parallel corpus and the NMT system is trained like normal, with no metadata to distinguish the two input languages.

Our proposed method proves to be an effective way of incorporating monolingual data into low-resource NMT. On Turkish↔English and Romanian↔English translation tasks, we see gains of up to 1.2 BLEU over a strong baseline with back-translation. Further analysis shows that the linguistic phenomena behind these gains are different from and largely orthogonal to back-translation, with our copied corpus method improving accuracy on named entities and other words that should remain identical between the source and target languages.

1 Introduction

Neural machine translation (NMT) systems require a large amount of training data to make generalizations, both on the source side (in order to interpret the text well enough to translate it) and on the target side (in order to produce fluent translations). This data typically comes in the form of parallel corpora, in which each sentence in the source language is matched to a translation in the target language. Recent work (Gulcehre et al., 2015; Sennrich et al., 2016b) has investigated incorporating monolingual training data (particularly on the target side) into NMT. This effectively converts machine translation into a semi-supervised problem that takes advantage of both labeled (parallel) and unlabeled (monolingual) data. Adding monolingual data to NMT is important because sufficient parallel data is unavailable for all but a few language pairs and domains.

In this paper, we introduce a straightforward method for adding target-side monolingual training data to an NMT system without changing its architecture or training algorithm. This method converts a monolingual corpus in the target language into a parallel corpus by copying it, so that each source sentence is identical to its corresponding target sentence. This copied corpus is then mixed with the original parallel data and used to train the NMT system, with no distinction made between the parallel and the copied data.

We focus on language pairs with small amounts of parallel data where monolingual data has the most impact. On the relatively low-resource language pairs of English↔Turkish and English↔Romanian, we find that our copying technique is effective both alone and combined with back-translation. This is the case even when no additional monolingual data is used (i.e. when the copied corpus and the back-translated corpus are identical on the target side). This implies that back-translation does not make full use of monolingual data in low-resource settings, which makes sense because it relies on low-resource (and therefore low-quality) translation in the reverse direction.

Proceedings of the Conference on Machine Translation (WMT), Volume 1: Research Papers, pages 148–156
Copenhagen, Denmark, September 711, 2017.

2 Related Work

Early work on incorporating monolingual data into NMT concentrated on target-side monolingual data. Jean et al. (2015) and Gulcehre et al. (2015) used a 5-gram language model and a recurrent neural network language model (RNNLM), respectively, to re-rank NMT outputs. Gulcehre et al. (2015) also integrated a pre-trained RNNLM into NMT by concatenating hidden states. Sennrich et al. (2016b) added monolingual target data directly to NMT using null source sentences and freezing encoder parameters while training with the monolingual data. Our method is similar, although instead of using a null source sentence, we use a copy of the target sentence and train the encoder parameters on the copied sentence.

Sennrich et al. (2016b) also created synthetic parallel data by translating target-language monolingual text into the source language. To perform this process, dubbed *back-translation*, they first trained an initial target→source machine translation system on the available parallel data. They then used this model to translate the monolingual corpus from the target language to the source language. The resulting back-translated data was combined with the original parallel data and used to train the final source→target NMT system. Since this back-translation method outperforms previous methods that only train the decoder (Gulcehre et al., 2015; Sennrich et al., 2016b), we use it as our baseline. In addition, our method stacks with back-translation in both the target→source and source→target systems; we can use source text to improve the back-translations and target text to improve the final outputs.

In the mirror image of back-translation, Zhang and Zong (2016) added source-side monolingual data to NMT by first translating the source data into the target language using an initial machine translation system and then using this translated data and the original parallel data to train their NMT system. Our method is orthogonal: it could improve the initial system or be used alongside the translated data in the final system. They also considered a multitask shared encoder setup where the monolingual source data is used in a sentence reordering task.

More recent approaches have used both source and target monolingual data while simultaneously training source→target and target→source NMT systems. Cheng et al. (2016) accomplished this by concatenating source→target and target→source NMT systems to create an autoencoder. Monolingual data was then introduced by adding an autoencoder objective. This can be interpreted as back-translation with joint training. He et al. (2016) similarly used a small amount of parallel data to pre-train source→target and target→source NMT systems; they then added monolingual data to the systems by translating a sentence from the monolingual corpus into the other language and then translating it back into the original language, using reinforcement learning with rewards based on the language model score of the translated sentence and the similarity of the reconstructed sentence to the original. Our approach also employs an autoencoder, but rather than concatenate two NMT systems, we have flattened them into one standard NMT system.

Our approach is related to multitask systems. Luong et al. (2016) proposed conjoined translation and autoencoder networks; we use a single shared encoder. Further work used the same encoder and decoder for multi-way translation (Johnson et al., 2016). We have repurposed the idea to inject monolingual text for low-resource NMT. Their work combined multiple translation directions (e.g. French→English, German→English, and English→German) into one system. Our work combines e.g. English→English and Turkish→English into one system for the purpose of improving Turkish→English quality. They used only parallel data; our goal is to inject monolingual data.

3 Neural Machine Translation

We evaluate our approach using sequence-to-sequence neural machine translation (Cho et al., 2014; Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) augmented with attention (Bahdanau et al., 2015). We briefly explain these models here.

Neural machine translation is an end-to-end approach to machine translation that learns to directly model $p(y \mid x)$ for a source-target sentence pair (x, y). The system consists of two recurrent neural networks (RNNs): the encoder and the decoder. In our experiments, the encoder is a bidirectional RNN with gated recurrent units (GRUs) that maps the source sentence into a vector representation. The decoder is an RNN language model conditioned on the source sentence. This is aug-

mented with an attention mechanism, which assigns weights to each of the words in the source sentence when modeling target words. This model is trained to minimize word-level cross-entropy loss; at test time, translations are generated using beam search.

4 Copied Monolingual Data for NMT

We propose a method for incorporating target-side monolingual data into low-resource NMT that does not rely heavily on the amount or quality of the parallel data. We first convert the target-side monolingual corpus into a bitext by making each source sentence identical to its target sentence; i.e., the source side of the bitext is a copy of the target side. We refer to this bitext as the *copied corpus*. The copied corpus is then mixed with the bilingual parallel corpus and no distinction is made between the two corpora. Finally, we train our NMT system with a single encoder and decoder using this mixed data. We are able to use the same encoder for both the parallel and the copied source sentences because we use byte pair encoding (Sennrich et al., 2016c) to represent the source and target words in the same vocabulary.

This copying method can also be combined with the back-translation method of Sennrich et al. (2016b). This is done by shuffling the parallel, back-translated, and copied corpora together into a single dataset and training the NMT system like normal, again making no distinction between the three corpora during training. We experiment with using the same monolingual data as the basis for both the back-translated and copied corpora (so that the target sides of the back-translated and copied corpora are identical) and with using two separate monolingual datasets for these purposes. Note that in the former case, each sentence in the original monolingual corpus occurs twice in the training data.

5 Experiments

5.1 Experimental Setup

5.1.1 Training Details

We train attentional sequence-to-sequence models (Bahdanau et al., 2015) implemented in Nematus (Sennrich et al., 2017). We use hidden layers of size 1024 and word embeddings of size 512. The models are trained using Adam (Kingma and Ba, 2015) with a minibatch size of 80 and a maximum sentence length of 50. We apply dropout (Gal and Ghahramani, 2016) in all of our EN↔TR and EN↔RO systems with a probability of 0.1 on word layers and 0.2 on all other layers. No dropout is used for EN↔DE. For all models, we use early stopping based on perplexity on the validation dataset. We decode using beam search on a single model with a beam size of 12, except for EN↔DE where we use a beam size of 5. For the experiments which use back-translated versions of the monolingual data, the target→source systems used to create the back-translations have the same setup as those used in the final source→target experiments.

Language pair	Parallel	Monolingual
EN↔TR	207 373	414 746
EN↔RO	608 320	608 320
EN↔DE	5 852 458	10 000 000

Table 1: Number of parallel and monolingual training sentences for each language pair.

5.1.2 Data and Preprocessing

We evaluate our models on three language pairs: English (EN) ↔ Turkish (TR), English ↔ Romanian (RO), and English ↔ German (DE). As shown in Table 1, these pairs each have vastly different amounts of parallel data. All of these languages have a substantial amount of monolingual data available.

The EN↔TR and EN↔DE data comes from the WMT17 news translation shared task,[1] while the EN↔RO data comes from the WMT16 shared task (Bojar et al., 2016). We use all of the available parallel data for each language pair, and the monolingual data comes from News Crawl 2015 (EN↔RO) or News Crawl 2016 (EN↔TR and EN↔DE). To create our monolingual datasets we randomly sample from the full monolingual sets.

For all language pairs, we tokenize and truecase the parallel and monolingual training data; we also apply byte pair encoding (BPE) to split words into subword units (Sennrich et al., 2016c). For each language pair, we learn a shared BPE model with 90,000 merge operations. Both the BPE model and the truecase model are learned on parallel data only (not on monolingual data). For RO→EN, we remove diacritics from the source training data, following the recommendation by Sennrich et al. (2016a).

[1] http://statmt.org/wmt17

	EN→TR		TR→EN		EN→RO	RO→EN	EN→DE		DE→EN	
BLEU	**2016**	**2017**	**2016**	**2017**	**2016**	**2016**	**2016**	**2017**	**2016**	**2017**
baseline	12.8	14.2	18.5	18.3	23.8	34.5	**33.3**	**26.6**	40.1	33.8
+ copied	**14.0**†	**15.2**†	**18.9**‡	**18.6**‡	**24.5**†	**35.7**†	**33.3**	26.3	**40.2**	**34.0**

Table 2: Translation performance in BLEU with and without copied monolingual data. Statistically significant differences are marked with † ($p < 0.01$) and ‡ ($p < 0.05$).

5.2 Translation Performance

We evaluate our models compared to a baseline containing parallel and back-translated data on the newstest2016 (all language pairs) and newstest2017 (EN↔TR and EN↔DE) test sets. For each model, we report case-sensitive detokenized BLEU (Papineni et al., 2002) calculated using `mteval-v13a.pl`.

The BLEU scores for each language pair and each system are shown in Table 2. The only difference between the baseline and the + *copied* systems is the addition of the copied corpus during training. Note that the copied and the back-translated corpora are created using identical monolingual data, which means that in the + *copied* system, each sentence from the monolingual corpus occurs twice in the training data (once as part of the copied corpus and once as part of the back-translated corpus).

For EN↔TR and EN↔DE, we use about twice as much monolingual as parallel data, so the ratio of parallel to back-translated to copied data is 1:2:2. For EN↔RO, we use a 1:1:1 ratio. In addition, for EN↔DE, we oversample the parallel corpus twice in order to balance the parallel and monolingual data.

For EN↔TR and EN↔RO, we observe statistically significant improvements (up to 1.2 BLEU) when adding the copied corpus. This indicates that our copied monolingual method can help improve NMT in cases where only a moderate amount of parallel data is available. For EN↔DE, we do not see improvements from adding the copied data; we conjecture that this occurs because this is a high-resource language pair. However, the EN↔DE systems trained with the copied corpus also do not perform any worse that those without.

5.3 Fluency

Adding copied target-side monolingual data results in a significant improvement in translation performance as measured by BLEU for EN↔TR and EN↔RO. Motivated by a desire to better understand the source of these improvements, we further experiment with the outputs for each system described in section 5.2. In particular, we want to examine whether these gains are simply due to the monolingual data improving the fluency of the NMT system.

In order to evaluate the fluency of each system, we train 5-gram language models for each language using KenLM (Heafield, 2011). The models are trained on the full monolingual News Crawl 2015 and 2016 datasets. This data is preprocessed as described in section 5.1, except that no subword segmentation is used.

We use these language models to measure perplexity on the outputs of the baseline systems (trained using parallel and back-translated data) and the + *copied* systems (trained using parallel, back-translated, and copied data). The language models are also queried on the reference translations for comparison. For all language pairs except EN↔RO, we concatenate newstest2016 and newstest2017 into a single dataset to find the perplexity.

Table 3 displays the perplexities for each system output and the reference. Interestingly, the perplexities for the baseline and the + *copied* systems are similar for all language pairs. In particular, improvements in BLEU (see Table 2) do not necessarily correlate to improvements in perplexity. This indicates that the gains from the + *copied* system may not solely be due to fluency.

5.4 Pass-through Accuracy

Since the copied monolingual data adds an autoencoder element to the NMT training, it is possible that the systems trained with copied data learn how to better pass through named entities and other relevant words than the baselines. In order to test this hypothesis, we detect words that are identical in each sentence in the source and the reference for the tokenized test data (excluding words that contain only one character and ignoring case). We then count how many of these words occur in the corresponding sentence in the translation output from each system. We calculate the pass-through

Perplexity	EN→TR	TR→EN	EN→RO	RO→EN	EN→DE	DE→EN
reference	700.0	146.7	202.4	118.1	231.0	116.5
baseline	**921.1**	**341.6**	**328.2**	248.4	**490.6**	317.3
+ copied	921.6	344.2	344.8	**245.5**	493.3	**314.2**

Table 3: Language model perplexities for the outputs of each NMT system.

Accuracy	EN→TR	TR→EN	EN→RO	RO→EN	EN→DE	DE→EN
baseline	77.3%	85.0%	71.5%	85.3%	78.5%	**91.4%**
+ copied	**82.0%**	**89.1%**	**78.5%**	**91.5%**	**78.6%**	91.1%

Table 4: Pass-through accuracy for the outputs of each NMT system.

accuracy as the percent of such words that appear in the output; these results are shown in Table 4.

For all language pairs except for EN↔DE, there is a large improvement in pass-through accuracy when the copied data is added during training. This closely mirrors the BLEU results discussed in section 5.2. These results suggest that a key advantage of using copied data is that the model learns to pass appropriate words through to the target output more successfully. Table 5 shows some examples of translations with improved pass-through accuracy for the + *copied* systems.

5.5 Additional EN-TR Experiments

In this section, we describe a number of additional experiments on EN→TR in order to investigate the effects of different experimental setups and aspects of the data. Note that the BLEU scores in this section are not directly comparable with those in Table 2, since a different subset of the monolingual data is used for some of these experiments. All BLEU scores reported in this section are on newstest2016 unless otherwise noted.

5.5.1 Double Back-Translated Data

In section 5.2, we report significant gains from our + *copied* systems over baselines trained on parallel and back-translated data for EN↔TR and EN↔RO, even while using the same monolingual data as the basis for both the copied and the back-translated corpora. However, in our experiments, we use particularly high-quality in-domain monolingual data. As a result, it is possible that these improvements are due to using this monolingual data twice (in the form of the back-translated and copied corpora) rather than to using the copied monolingual corpus.

In order to evaluate this, we consider an additional configuration in which we train using two copies of the same back-translated corpus (instead of using one copy of each of the back-translated corpus and the copied corpus). The results for this experiment are in Table 6. For both test sets, the + *copied* system performs better than the system with double back-translated data by about 1 BLEU point. This indicates that our copied corpus improves NMT performance, and that this is not simply due to the higher weight given to the high-quality monolingual data.

5.5.2 Different Copied Data

In our initial experiments, we use the same monolingual corpus to create the back-translated and the copied data. Here, we consider a variation in which we use different monolingual data for these purposes. This is done by cutting the monolingual corpus in half and back-translating only half of it, leaving the rest for copied data. Note that this means that the original monolingual corpus is the same size (twice the size of the parallel data; see Table 1), but each monolingual sentence only occurs once in the training data, rather than twice as before.

The results for these experiments are shown in Table 7. The baseline is trained on back-translations of all of the monolingual data, and the + *same copied* system contains the full copied corpus. The + *different copied* system uses different data for copying and back-translation. Both copied systems outperform the baseline, although the + *same copied* system does slightly better.

5.5.3 Copied Data Without Back-translation

Our results in section 5.2 show that our copied corpus method stacks with back-translation to improve translation performance when there is not much parallel data available. In this section, we study whether the copied corpus can aid NMT when no back-translated data is used. If so, this would be advantageous, as the copied corpus method is much simpler to apply than back-

	RO→EN
source	... a afirmat Angel Ubide, analist șef în cadrul **Peterson** Institute for International Economics.
reference	... said Angel Ubide, senior fellow at the **Peterson** Institute for International Economics.
baseline	... "said Angel Ubide, chief analyst at the **Carson** Institute for International Economics.
+ copied	... "said Angel Ubide, chief analyst at **Peterson** Institute for International Economics.
source	Les **Dissonances** a aparut pe scena muzicala în 2004 ...
reference	Les **Dissonances** appeared on the music scene in 2004 ...
baseline	Les **Dissonville** appeared on the music scene in 2004 ...
+ copied	Les **Dissonances** appeared on the music scene in 2004 ...
	TR→EN
source	**Metcash**, Bay **Douglass**'ın yorumlarına bir yanıt vermeyi reddetti.
reference	**Metcash** has declined to respond publicly to Mr **Douglass**' comments.
baseline	**Metah** declined to give an answer to Mr. **Doug**'s comments.
+ copied	**Metcash** declined to respond to a response to Mr. **Douglass**'s comments.
source	PSV teknik direktörü Phillip **Cocu**, şöyle dedi: "Çok kötü bir sakatlanma."
reference	Phillip **Cocu**, the PSV coach, said: "It's a very bad injury."
baseline	PSV coach Phillip **Coker** said: "It was a very bad injury.
+ copied	PSV coach Phillip **Cocu** said: "It's a very bad injury."

Table 5: Comparison of translations generated by baseline and + *copied* systems.

BLEU	**2016**	**2017**
parallel + back-translated	12.4	14.2
parallel + double back-translated	13.1	14.1
parallel + back-translated + copied	**14.0**	**15.2**

Table 6: EN→TR translation performance when using the back-translated corpus twice vs. the back-translated and copied corpora.

	BLEU
baseline	12.4
+ same copied	**13.6**
+ different copied	13.3

Table 7: EN→TR translation performance when using the same or different data for copied and back-translated corpora.

translation and does not require the training of an additional target→source machine translation system. We experiment with both a small copied corpus (about 200k sentences) and a large copied corpus (about 400k sentences).

The results for systems trained with only parallel and copied data are in Table 8. Both the small copied corpus and the large copied corpus yield large improvements (2.3-2.6 BLEU) over using parallel data only, and their performance is only slightly worse (0.3-0.4 BLEU) than the corresponding systems trained with only back-translated and parallel data.

5.5.4 Source Monolingual Data

Although we have concentrated thus far on incorporating target-side monolingual data into NMT, source-side monolingual data also has the potential to help translation performance. In particular, a source copied corpus can be used when training the target→source system for back-translation. Here, we test this strategy on EN→TR NMT with EN monolingual data. For this purpose, we randomly sample about 400k English sentences (twice the size of the parallel corpus) from the News Crawl 2015 monolingual corpus.

The results for this experiment are shown in Table 9. Although both copied systems improve over the baseline, adding the EN monolingual data does not result in further improvement over the target-only copied model, despite taking much longer to train.

	BLEU
parallel only	9.4
parallel + small copied	11.7
parallel + large copied	12.0
parallel + small back-translated	**12.0**
parallel + large back-translated	**12.4**

Table 8: EN→TR translation performance without back-translated data. We include systems trained with parallel and back-translated data (without copied data) for comparison.

	BLEU
baseline	12.4
+ copied	**13.6**
+ EN data	**13.6**

Table 9: EN→TR translation performance with EN monolingual data.

BLEU	1:1	2:1	3:1
baseline	12.0	12.4	12.8
+ copied	**13.0**	**13.6**	**13.8**

Table 10: EN→TR translation performance with different amounts of monolingual data.

5.5.5 Amount of Monolingual Data

Finally, we study the effectiveness of the copied monolingual corpus when the amount of monolingual data is varied. We consider three different monolingual corpus sizes: the same size as the parallel data (200k sentences; *1:1*), twice the size of the parallel data (400k sentences; *2:1*), and three times the size of the parallel data (600k sentences; *3:1*). We compare these different sizes for the baseline (parallel and back-translated data) and the + *copied* systems (parallel, back-translated, and copied data, where the back-translated and copied data are identical on the target side). Each smaller monolingual corpus is a subset of the larger monolingual corpora. Note that we do not oversample the parallel data to balance the different data sources.

Table 10 displays the results when different amounts of monolingual data are used. Note that we vary the amount of back-translated data in the baseline and of back-translated and copied data in the + *copied* system. For both the baseline and + *copied*, adding more monolingual data consistently yields small improvements (0.2-0.6 BLEU). In addition, the + *copied* system performs about 1.0 BLEU better than the baseline regardless of the amount of monolingual data. This is surprising since we do not oversample the parallel data at all. For the *2:1* and *3:1* cases, the systems see far less parallel than synthetic data, but the overall translation performances still improve.

6 Discussion

Our proposed method of using a copied target-side monolingual corpus to augment training data for NMT proved to be beneficial for EN↔TR and EN↔RO translation, resulting in improvements of up to 1.2 BLEU over a strong baseline. We showed that our method stacks with the previously proposed back-translation method of Sennrich et al. (2016b) for these language pairs. For EN↔DE, however, there was no significant difference between systems trained with the copied corpus and those trained without it. There was much more parallel training data for EN↔DE than for EN↔RO (nearly 10 times as much) and EN↔TR (about 28 times as much), so it is possible that the gains that would have come from the copied corpus were already achieved with the parallel data. Overall, the copied monolingual corpus either helped or was indifferent, so training with this corpus is not risky. In addition, it does not require any more monolingual data besides what is used for back-translation.

We initially assumed that the copied monolingual corpus was helping to improve the fluency of the target outputs. However, further study of the outputs did not necessarily support this assumption, as noted in section 5.3. Our method did improve accuracy when copying proper nouns and other words that are identical in the source and target languages; this is at least part of the explanation for the increases in BLEU score when using the copied corpus.

Subsequent experiments revealed various factors that influenced the effectiveness of the copied monolingual corpus. An unexpected finding was that doubling and tripling the size of the monolingual corpus (whether used as copied or back-translated data) resulted in small improvements (0.2-0.6 BLEU). We had originally thought that using much more monolingual than parallel data would result in a worse performance, since the system would see true parallel data less often than copied or back-translated data, but this did not turn out to be the case. Not having to limit the amount of monolingual data based on the availability of parallel data is an advantage for language pairs with much more monolingual than parallel data.

7 Conclusion

In this paper, we introduced a method for improving neural machine translation using monolingual data, particularly for low-resource scenarios. Augmenting the training data with monolingual data in which the source side is a copy of the target side proved to be an effective way of improving EN↔TR and EN↔RO translation, while not damaging EN↔DE (high-resource) translation. This technique could be used in combination with back-translation or with parallel data only. In addition, using much more monolingual than parallel data did not hinder performance, which is beneficial for the common case where a large amount of monolingual data is available but the language pair has little parallel data.

In the future, we plan on studying the effects of the quality of the monolingual data, since our copied corpus technique might in principle pose the risk of adding noise to the NMT system. In particular, we would like to apply a data selection method when creating the monolingual corpus, as the similarity of the monolingual and parallel data has been shown to have an effect on NMT (Cheng et al., 2016). We also hope to find an effective way of adding source monolingual training data. Finally, it would be interesting to do a manual evaluation of our method to confirm the BLEU and perplexity findings reported in sections 5.2 and 5.3.

Acknowledgments


This work was conducted within the scope of the Horizon 2020 Innovation Action *Health in My Language*, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402. This work was partially funded by the Amazon Academic Research Awards program. We used Azure credits donated by Microsoft to The Alan Turing Institute. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198. Association for Computational Linguistics.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 1965–1974. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29*.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29*.

Kenneth Heafield. 2011. KenLM: faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal neural machine translation systems for WMT15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

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. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *4th International Conference on Learning Representations*.

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 ACL*, pages 311–318. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin

Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a toolkit for neural machine translation. In *Proceedings of the EACL 2017 Software Demonstrations*, pages 65–68. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of NAACL-HLT*, pages 86–96. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*.

Jiajun Zhang and Chengqing Zong. 2016. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545. Association for Computational Linguistics.