# Using Multiple Subwords to Improve English-Esperanto Automated Literary Translation Quality

**Alberto Poncelas[1,2], Jan Buts[1,2], James Hadley[1,2], Andy Way[2]**

[1]Trinity Centre for Literary and Cultural Translation, Trinity College Dublin, Ireland
`butsj@tcd.ie,hadleyj@tcd.ie`
[2]ADAPT Centre, School of Computing, Dublin City University, Ireland
`name.surname@adaptcentre.ie`

## Abstract

Building Machine Translation (MT) systems for low-resource languages remains challenging. For many language pairs, parallel data are not widely available, and in such cases MT models do not achieve results comparable to those seen with high-resource languages.

When data are scarce, it is of paramount importance to make optimal use of the limited material available. To that end, in this paper we propose employing the same parallel sentences multiple times, only changing the way the words are split each time. For this purpose we use several Byte Pair Encoding models, with various merge operations used in their configuration.

In our experiments, we use this technique to expand the available data and improve an MT system involving a low-resource language pair, namely English-Esperanto.

As an additional contribution, we made available a set of English-Esperanto parallel data in the literary domain.

## 1 Introduction

In this paper, we use the constructed language Esperanto to illustrate potential improvements in the automatic translation of material from low-resource languages. Languages are considered low-resource when there is little textual material available in the form of electronically stored corpora. They pose significant challenges in the field of Machine Translation (MT), since it is difficult to build models that perform adequately using small amounts of data.

Multiple techniques have been developed to improve MT in conditions of data scarcity. A popular approach is to translate indirectly via a pivot language (Utiyama and Isahara, 2007; Firat et al., 2017; Liu et al., 2018; Poncelas et al., 2020a). Moreover, indirect translation can be used

for creating additional training data. A further useful technique for expanding the dataset is back-translation (Sennrich et al., 2016a). This procedure consists of automatically translating a monolingual text from the target language into the selected source language, and then using the resulting parallel set as training data so the model benefits from this additional information. Although the quality of these sentence pairs is not as high as that of human-translated sentences (the source side contains mistakes produced by the MT system), the pairs are still useful when used as training data, because they do often improve the models (Poncelas et al., 2019a).

Nonetheless, for some languages, the available data are in such short supply that MT models used for generating back-translated sentences may produce a high proportion of noisy sentences. The use of noisy sentences for building MT models could ultimately have a negative impact on the quality of the MT system's outputs (Goutte et al., 2012), and therefore they are often removed (Khadivi and Ney, 2005; Taghipour et al., 2010; Popović and Poncelas, 2020).

We propose employing another technique to augment datasets: using the same set of sentences multiple times, but in slightly altered form each time. Specifically, we modify the sentences by using different Byte Pair Encoding (BPE) (Sennrich et al., 2016b) merge operations. We perform a fine-grained analysis, exploring the use of different splitting options on the source side, on the target side, and on both sides.

## 2 Previous work

This research is inspired by techniques for augmenting the training set artificially. One of these techniques is back-translation (Sennrich et al., 2016a), which involves creating artificial source-side sen-

tences by translating a monolingual set in the target language. Similar techniques include the use of several models to generate sentences (Poncelas et al., 2019b; Soto et al., 2020), or the use of synthetic data on the target side (Chinea-Rios et al., 2017; Li et al., 2020).

A technique that involves multiple segmentation is subword regularization (Kudo, 2018), in which candidate sentences with different splits are sampled, either probabilistically or using a language model for training.

In the work of Poncelas et al. (2020b), different splits are used to build an English-Thai MT model. As the Thai language does not use whitespace separation between words, different splits can be applied, to address the fact that all the words and sub-words are joined together in the final output.

More recently, Provilkov et al. (2020) introduced BPE-dropout, an improvement on standard BPE consisting of randomly dropping merges when training the model, such that a single word can have several segmentations.

## 3   The Esperanto language

This article is concerned with improving MT models for Esperanto, the most successful constructed international language (Blanke, 2009). It was created in the late nineteenth century, and is said to be currently spoken by over 2 million people, spread across more than 100 countries (Eberhard et al., 2020). During its first century of development, Esperanto was principally maintained by means of membership-based organisations. Currently, internet applications such as Duolingo are supporting the wider spread of the language among new enthusiasts. While many Esperanto speakers have sought to develop the language through translation, the body of work available - particularly in digital formats - remains relatively small, making Esperanto a clear example of a low-resource language.

Esperanto loosely derives its lexicon from several Indo-European languages, and shares some typological characteristics with, among others, Russian, English, and French (Parkvall, 2010). In contrast to most natural languages, Esperanto's most distinctive characteristic is its regularity. The grammar consists of a very limited set of operations, to which there are, in principle, no exceptions. Furthermore, the language is agglutinative, and its suffixes are independently meaningful and invariable. For instance, *virino*, the word for "woman", con-

sists of the compound parts *vir* [adult human], *in* [female], and *o* [entity] (as the 'o' ending is used for all nouns). The word for "mother", *patrino*, largely refers to the same semantic categories, and is therefore structurally highly similar.

As a consequence of this internal consistency, Esperanto learners can quickly expand their vocabulary by learning to segment words into their various parts, which can then be used to construct new words by morphological analogy. Because of its affinity with many other languages, and because of the thoroughly logical composition of its vocabulary, Esperanto has historically been central to several experiments in MT, most notably regarding its potential function as a pivot language between European languages (Gobbo, 2015). In this study, however, we focus on automatic translation into Esperanto for its own sake.

## 4   Research Questions

We propose building MT models using training data composed of a dataset split into multiple variants with a different configuration of BPE, as presented in Figure 1. At the top of the figure, one can see that the same parallel set has been processed using BPE with 89,500, 50,000 and 10,000 operations (trained separately for each language). The MT model represented on the left has been built using the same dataset replicated three times, the only difference being that on the target side, different splits were implemented. Similarly, the MT model in the centre is built with different splits on the source side. The last model, represented on the right, combines different splits both on the source and the target side.

In order to evaluate the models, we use a test set that is split with a single BPE strategy (i.e. using 89,500 merge operations, the default proposed in the work of Sennrich et al. (2016b)). Therefore, using different merge operations on the source side of the training data may not have as big an impact as when they are applied to the target side (not all the words will match those in the test set). However, the addition of other BPE configurations could in principle still be useful to improve modeling for the source language.

In Section 5 we describe the settings of the MT and the data used for training. In Section 6 we analyze the results achieved by the baseline system.

This paper's experiments are divided into three sections. Each of these sections describes and also

provides the evaluation of a model. The sections are the following:

- Combination of dataset with different merge operations on the target side (Section 7.1).

- Combination of dataset with different merge operations on the source side (Section 7.2).

- Combination of dataset with different merge operations on both the source and target side (Section 7.3).

In Section 8, we compare translation examples from the different models and analyze the different outcomes.

Finally, in Section 9 we conclude and propose how these experiments could be expanded in future work.

## 5 Experimental Settings

The NMT systems we build are Transformer (Vaswani et al., 2017) models, based on OpenNMT (Klein et al., 2017). Models are trained for a maximum of 30K steps using the recommended parameters.[1] We have selected the model with the lowest perplexity on the development set.

### 5.1 Dataset

For training the models we use the *Tatoeba*, *GlobalVoices* and *bible-uedin* (Christodouloupoulos and Steedman, 2015) datasets from OPUS project.[2] Our dataset thus contains material from the Bible, from news sources, and from less domain-specific multilingual translation examples. The sentences are randomly shuffled, after which 302,768 sentences are used as a training set and the other 1,000 as our dev set. All the sentences are tokenized and truecased.

BPE is applied using several merge operations. We use 89,500 operations as a starting point and explore other splits that produce smaller subword units (by using a lower number of merge operations). In our experiments we work with 50,000, 20,000 and 10,000 operations.

We also concatenate the dev set using the same configuration of BPE.

### 5.2 Test Set

In order to evaluate the quality of the models, two test sets are translated. The test sets are the same for all models. In addition to tokenization and truecase, we also use BPE with 89,500 merge operations. We do not use (or combine) other BPE configurations. The translations are evaluated using the BLEU (Papineni et al., 2002) metric.

The first test set is taken from the OPUS (*Books*) dataset (Tiedemann, 2012) (1562 sentences). Specifically, the test set consists of material from two texts available in English and in Esperanto translation, namely Carroll's *Alice's Adventures in Wonderland* (Carroll and Kearney, 1865 (1910) and Poe's *The Fall of the House of Usher* (Poe and Grobe, 1839 (2000).[3]

The second test set (which contains 1256 sentences) consists of an English and an Esperanto version of Oscar Wilde's *Salomé* (Wilde et al., 1891 (1894, 1910), [4] a play originally written in French. As an additional contribution to this paper, we have made a set of aligned sentences from the texts available via OPUS.[5]

Both test sets are in the literary domain, which is especially challenging (Toral and Way, 2018) for MT models. Not only do the test sets contain numerous personal names and uncommon vocabulary, they are also highly creative and, at times, experimental. For instance, in *Alice's Adventures in Wonderland*, grammatical and lexical principles are often challenged on purpose to portray a character's individual traits (i.e. the Mock Turtle sings of *Beau–ootiful soo–oop!*. In *Salome*, characters regularly produce complex similes and metaphors to describe one another. The text is a variation on a religious theme, and heavily draws on Biblical imagery. While such material is highly challenging, the inclusion of Biblical matter in the training data may have a positive impact on the overall results.

## 6 Baseline MT

In Table 1 we present the models trained with the training data using different merge operations on the target side.

The rows of the table correspond to the evaluation of the model, using the same data. The only
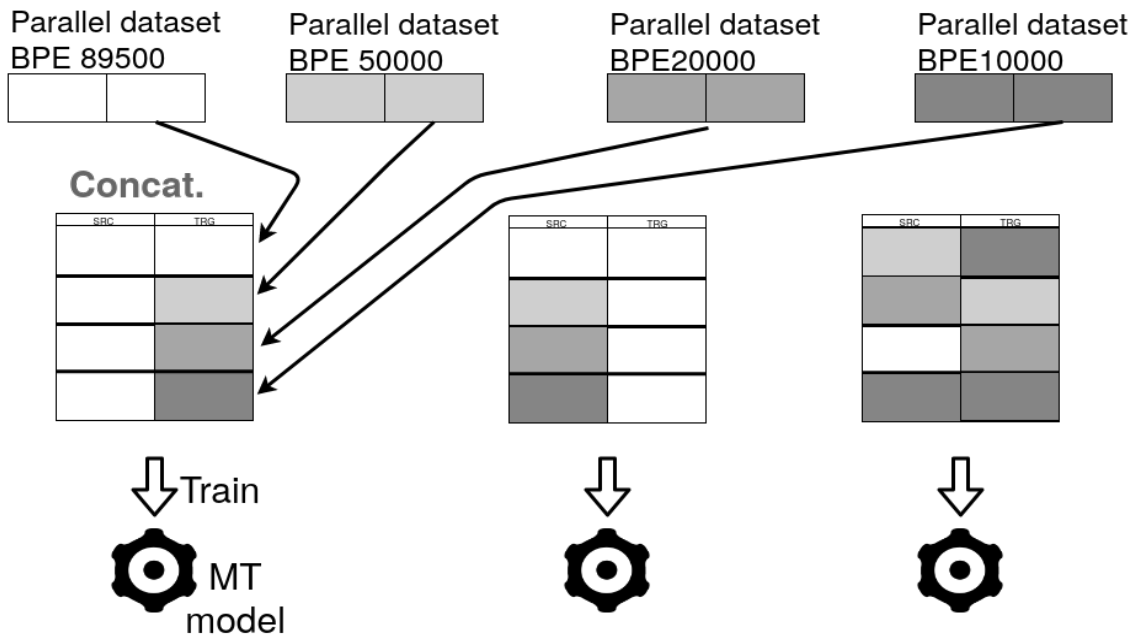
Figure 1: Diagram with the experiments

| Traindata | Books | Salome |
|-----------|-------|--------|
| TRG89500 | 6.72 | 18.03 |
| TRG50000 | 6.42 | 18.74 |
| TRG20000 | 6.89 | 18.49 |
| TRG10000 | 6.65 | 18.41 |

Table 1: BLEU scores of the **Books** and **Salome** test sets when translated using the Baseline MT.

difference is the number of BPE merge operations that have been used on the target side.

As the test set is split using 89,500 merge operations, it would not be beneficial to apply BPE with merge operations other than 89,500 on the source side. In fact, when using BPE with 50,000, 20,000 and 10,000 operations on the source side, the BLEU score for the translation of the Books data is only 5.75, 5.70 and 5.76, respectively, and 14.30, 13.24, and 14.53 for *Salome*

Table 1 shows that the four models achieve similar results. As mentioned before, the *Books* set contains complex grammatical and lexical constructions, which makes it more difficult to translate. This is also evidenced in the table as BLEU scores of the *Books* set are lower than those of the *Salome* set. Moreover, there is no correlation between the number of merge operations and the performance. For example, we observe a small drop in the performance when decreasing the number of merge operations from 89,500 to 50,000, but the performance

improves slightly when the number of operations is further decreased to 20,000.

## 7 Experiments

### 7.1 Different Merge Operations on the Target Side

In the first set of experiments we explore the models when the sentences in the parallel set are replicated by changing only the number of BPE merge operations used on the target side. We perform two sets of experiments: one where we keep the duplicates (sentences that remain the same after being split with different BPE configurations), and another where duplicates are removed.

In Table 2 we present the results of the models when trained with a different concatenation of datasets. The first column specifies the datasets used in the training. For example, the row *TRG89500 & TRG50000* indicates that the training set used for building the MT model consists of sentences split using 89,500 and 50,000 merge operations, respectively

We mark in bold those scores that exceed 6.89 BLEU points, i.e. the maximum score achieved by the baseline models presented in Table 1. The scores receive an asterisk when the improvements are statistically significant at p=0.01. Statistical significance has been computed using Bootstrap Resampling (Koehn, 2004).

In the table, we find that scores tend to be higher

| | Traindata | Books | Salome |
|---|---|---|---|
| keep duplicates | TRG89500 & TRG50000 | 6.60 | 18.25 |
| | TRG89500 & TRG20000 | 6.89 | 18.49 |
| | TRG89500 & TRG10000 | **7.79*** | **19.41*** |
| | TRG50000 & TRG20000 | **7.17** | 17.96 |
| | TRG50000 & TRG10000 | **6.94** | **18.91** |
| | TRG20000 & TRG10000 | **7.48*** | **19.53*** |
| | TRG89500 & TRG50000 & TRG20000 | 6.64 | 18.63 |
| | TRG89500 & TRG50000 & TRG10000 | **7.38** | 18.64 |
| | TRG89500 & TRG20000 & TRG10000 | **7.91*** | **19.29** |
| | TRG50000 & TRG20000 & TRG10000 | **7.51*** | **18.86** |
| | TRG89500 & TRG5000 & TRG20000 & TRG10000 | **7.25** | 18.68 |
| remove duplicates | TRG89500 & TRG50000 | **7.47*** | **18.51** |
| | TRG89500 & TRG20000 | **7.56*** | **19.08** |
| | TRG89500 & TRG10000 | **7.98*** | **18.78** |
| | TRG50000 & TRG20000 | **6.93** | 18.47 |
| | TRG50000 & TRG10000 | **7.54*** | **18.62** |
| | TRG20000 & TRG10000 | **7.79*** | 18.32 |
| | TRG89500 & TRG50000 & TRG20000 | **7.22** | 18.40 |
| | TRG89500 & TRG50000 & TRG10000 | **7.53*** | **18.99** |
| | TRG89500 & TRG20000 & TRG10000 | **7.90*** | 18.76 |
| | TRG50000 & TRG20000 & TRG10000 | **7.39** | **19.07** |
| | TRG89500 & TRG5000 & TRG20000 & TRG10000 | **7.75*** | 18.84 |

Table 2: Model performance using different merge operations on the target side.

when duplicate sentence pairs are removed. By doing this the dataset is reduced by between 30% and 45%. In the second subtable, all the BLEU scores indicate improvements over the baseline, whereas in the first subtable some models, such as *TRG89500 & TRG50000*, have a lower score.

The best performance is seen when for the multiple settings used, the number of merge operations differs greatly. For example, the highest scores are achieved when mixing 89,500 and 10,000 operations (i.e. the *TRG89500 & TRG10000* rows in both subtables), the uppermost and the lowermost number of operations used in the experiments. The same principle holds true for those models built by combining three or four datasets.

## 7.2 Different Merge Operations on the Source Side

The next set of experiments explores the use of several merge operations on the source side. In this case, when combining the datasets, we ensure that the SRC89500 set is used, as the test set has been processed using 89,500 operations. We present the results in Table 3. Those scores that are higher than the baselines of Table 1 are marked in bold.

Our observations are similar to those obtained in Section 7.1. The best results are observed when the duplicate sentences are removed (between 25% and 40% of the sentences are removed) and the merge operation settings are the furthest apart (89,500 and 10,000).

Most of the models using several BPE configurations on the source side perform better than the baseline models. However, when compared to the experiments in the previous section (Table 2), the performance is lower.

## 7.3 Different Merge Operations on both Source and Target Side

The last set of experiments consists of building a model with data created using different splits both on the source and on the target side.

We perform experiments based on the outcomes observed in the previous section. Thus, two models are built. One combines the datasets split using BPE with 89,500 and 10,000 merge operations (both source and target side) and the other model, *All*, combines the dataset with all the splits (i.e.

| | Traindata | Books | Salome |
|---|---|---|---|
| keep duplicates | SRC89500 & SRC50000 | **7.19** | **19.07** |
| | SRC89500 & SRC20000 | **7.06** | 18.46 |
| | SRC89500 & SRC10000 | **7.21** | **19.28** |
| | SRC89500 & SRC50000 & SRC20000 | **7.25** | **19.16** |
| | SRC89500 & SRC50000 & SRC10000 | **7.09** | **19.15** |
| | SRC89500 & SRC20000 & SRC10000 | **6.96** | **18.98** |
| | SRC89500 & SRC50000 & SRC20000 & SRC10000 | 6.79 | 18.08 |
| remove duplicates | SRC89500 & SRC50000 | 6.44 | **18.54** |
| | SRC89500 & SRC20000 | **7.05** | **18.86** |
| | SRC89500 & SRC10000 | **7.43*** | **19.35*** |
| | SRC89500 & SRC50000 & SRC20000 | **7.42*** | 18.63 |
| | SRC89500 & SRC50000 & SRC10000 | 6.66 | **19.37** |
| | SRC89500 & SRC20000 & SRC10000 | **7.22** | **19.40*** |
| | SRC89500 & SRC50000 & SRC20000 & SRC10000 | 7.31 | 19.08 |

Table 3: Model performance using different merge operations on the source side.

| Traindata | Books | Salome |
|---|---|---|
| SRC89500 & SRC10000 & TRG89500 & TRG50000 | **7.99*** | 17.78 |
| All | **8.11*** | **19.70*** |

Table 4: Model performance using different merge operations both on the source and target side.

89,500, 50,000, 20,000 and 10,000).[6] The duplicates are removed, as this approach showed the best results.

We present the translation quality of the test set using these models in Table 4. We see that the use of different splits on both the source and target sides tends to achieve the best results when compared both to baselines and to the experiments in the previous sections[7].

# 8 Comparison of Outputs

In Table 5, we show some translation examples of the models that, as discussed in the previous sections, achieved the best performance. We mark in bold some important differences across the translations.

The first example, drawn from *Alice in Wonderland*, contains a joke. Alice, who is collecting her thoughts, aims to voice her opinion, and starts out by saying *I don't think....* Before she can finish her

sentence, however, the Mad Hatter interrupts her by stating that in that case, she should not speak. The human Esperanto translation makes this joke very explicit by repeating the emphasis on 'not thinking', whereas in English the transition is more subtle. Two of the systems, while differing in exact word order, succeed in reproducing the joke (TRG89500 and TRG89500 & TRG10000). In the other two models, either the crucial element *do* [so], which realises the inference, is omitted, or the meaning is mistakenly changed to a positive imperative: *vi devus diri* [you should say]. It can further be observed in the sentences that none of the systems translates the Hatter's name meaningfully. Either the name remains the same, or it is slightly altered from the original, in a seemingly random manner. Interestingly, Alice's name is adapted to Alico, which conforms to the rule that all Esperanto names end in *-o* (or, in some cases *-a*), but the adaptation does not equal the human choice for *Alicio*.

The second example, also taken from *Alice's Adventures in Wonderland*, is concerned with a particular fixed expression in the English language: *venture to say*. The baseline system does not translate this mark of politeness, while the other models do provide varying translations (i.e. *decidis*, *sukcesis* and *entrepenis*, which correspond to the past tenses of the verb *to decide*, *to succeed* and *to undertake*). While none of them is completely correct (when compared to the human translation), all of them are fairly transparent in context, and foreground different aspects of meaning contained in

---

[6]Note that we use all possible combinations. For example, the training set of the *All* model is built combining $4 * 4 = 16$ datasets.

[7]We observed that the output tends to be more similar to the splits following the TRG89500 configuration.

| system | sentence |
|---|---|
| source | said Alice, very much confused, "I don't think–" |
| | "Then you shouldn't talk," said the Hatter . |
| reference | Alicio , tre konfuzite, respondis... : "mi ne pensas-" |
| | "se vi ne pensas, vi ne rajtas paroli," diris la Ĉapelisto . |
| TRG89500 | Alico, tre konfuzita; mi ne pensas . |
| | "**do vi ne** parolu," diris la Hater. |
| SRC89500 & SRC10000 | Alico... "**vi ne** parolu," diris la Hatar. |
| TRG89500 & TRG10000 | "**vi do ne** parolu," diris la Hatter. |
| All | "**vi devus** diri, " diris la Hater. |
| source | but she did not venture to say it out loud . |
| reference | sed tion ŝi ne kuraĝis diri laŭte . |
| TRG89500 | sed ŝi ne **diris** tion laŭte . |
| SRC89500 & SRC10000 | sed ŝi ne **decidis diri** tion laŭte. |
| TRG89500 & TRG10000 | sed ŝi ne **intence diri** tion. |
| All | sed ŝi ne **entreprenis diri** tion laŭte. |
| source | it is like a knot of serpents coiled round thy neck. |
| reference | ili similas al fasko da nigraj serpentoj, kiun oni ĵetis ĉirkaŭ vian kolon . |
| TRG89500 | ĝi similas **nodon de turmentoj** ĉirkaŭ via kolo. |
| SRC89500 & SRC10000 | via kolo estas kiel **nodo streĉata**. |
| TRG89500 & TRG10000 | estas kiel **nodo da serpento**j ĉirkaŭ via kolo. |
| All | ĝi estas kiel **nodo de serpentoj** ĉirkaŭ via kolo. |
| source | it is like a pomegranate cut in twain with a knife of ivory. |
| reference | kiel granato, tranĉita per ebura tranĉilo . |
| TRG89500 | kiel granato, kiu **falas en du** pecojn de eburo . |
| SRC89500 & SRC10000 | kiel granato, kiu **tranĉiĝis en du** per tranĉilo de eburo. |
| TRG89500 & TRG10000 | ĝi similas al granato, **tranĉita en du** kun tranĉilo de eburo. |
| All | kiel granato **eltranĉita en du** kun tranĉilo ebura. |

Table 5: Translation examples from the test set.

the English *venture*.

With reference to the *Salome* test set, we find in the entire translated text numerous small and relatively inconsequential vocabulary differences across systems (e.g. *veston* or *mantelon* for referring to a piece of clothing), as well as varying preferences for orthographically similar verb tenses (e.g *lacigis* or *lacigas*, past and present tense of the verb *to tire* or *wear out*). At times, the systems differ in their translation of multi-word units such as *sacred person*, which is translated either as the literal *sankta homo* or as the more interpretative *sanktulo* [saint]. Overall, the systems perform well when translating the play's dense symbolism, as illustrated in Table 5.

The examples in the table are similes, which start with the explicit comparative phrase *it is like*. In the first example, the baseline system does not manage to reproduce the reference to *serpentoj* [snakes], although the mention of *turmentoj* [afflictions] does offer an interesting metaphorical perspective. The system *SRC89500 & SRC10000* does not produce a correct translation, but those systems trained with different splits on the target side (i.e. the *SRC10000 & SRC89500* and *All* systems) provide a remarkably good translation of the source. Similarly, in the last example included in the table, the baseline system fails to reproduce the meaning of the original (the knife falls apart instead of cutting the fruit), whereas all systems with multiple segmentation are successful in conveying a variant of the poetic image presented in the source text.

In short, the examples in Table 5 indicate that a combination of different merge operations may improve results for translation into Esperanto, a language for which limited resources are available. In a number of cases, the systems succeed in translating highly uncommon constructions in the context of humorous and poetic literary discourse.

## 9  Conclusion and Future Work

In this work, we have aimed to improve an English-Esperanto MT system by using multiple instances of the same sentence pair, split with different configurations of BPE.

In our experiments, the best performance tends to be achieved when splitting strategies are applied both on the source and target side, duplicate parallel sentences are removed, and the number of merge operations used are very different from each other. In our experiments, the best results are achieved

when all the split-combinations are used on both sides.

Although the goal of these experiments is to find a technique to improve the MT models when the available data are very limited, this technique could also be applied in scenarios where data are abundant. It should be noted that Esperanto is perhaps a particularly suitable candidate for word-split methods, as the language's vocabulary consists of fixed chunks that are combined to form transparent compounds. However, the techniques applied here are in principle language-independent.

Finally, although we demonstrated that combining sentences with different merge operations improves the model, in this paper we could not determine the best configuration to use. Similarly, the test set that we used was processed using 89,500 merge operations. If the test set had been processed with a different BPE configuration the performance could have been different, especially when using models with different split configurations on the source side. Extensions of this work could involve finding an optimal configuration for achieving the best results, or testing the performance when combined with other word-splitting techniques.

## Acknowledgments

## References

Detlev Blanke. 2009. Causes of the relative success of esperanto. *Language Problems and Language Planning*, 33(3):251–266.

Lewis Carroll and E.L. Kearney, trans. 1865 (1910). *Alice's adventures in wonderland (La aventuroj de Alicio en Mirlando)*. Project Gutenberg.

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

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

David M Eberhard, Gary F Simons, and Charles D Fennig. 2020. *Ethnologue: Languages of the World*, twenty-third edition. SIL International, Dallas, TX, USA.

Orhan Firat, Kyunghyun Cho, Baskaran Sankaran, Fatos T Yarman Vural, and Yoshua Bengio. 2017. Multi-way, multilingual neural machine translation. *Computer Speech & Language*, 45:236–252.

Federico Gobbo. 2015. Machine translation as a complex system, and the phenomenon of esperanto. *Interdisciplinary Description of Complex Systems*, 13(2):264–274.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of Association for Machine Translation in the Americas, AMTA*, San Diego, USA.

Shahram Khadivi and Hermann Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *International Conference on Application of Natural Language to Information Systems*, pages 263–274.

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia.

Hongzheng Li, Jiu Sha, and Can Shi. 2020. Revisiting Back-Translation for Low-Resource Machine Translation Between Chinese and Vietnamese. *IEEE Access*, 8:119931–119939.

Chao-Hong Liu, Catarina Cruz Silva, Longyue Wang, and Andy Way. 2018. Pivot machine translation using chinese as pivot language. In *China Workshop on Machine Translation*, pages 74–85, Wuyishan, China.

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

Mikael Parkvall. 2010. How european is esperanto?: A typological study. *Language Problems and Language Planning*, 34(1):63–79.

Edgar Allan Poe and Edwin Grobe, trans. 1839 (2000). *The fall of the house of usher (La Falo De Uŝero-Domo)*. Project Gutenberg.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019a. Adaptation of machine translation models with back-translated data using transductive data selection methods. In *20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Alberto Poncelas, Pintu Lohar, Andy Way, and James Hadley. 2020a. The Impact of Indirect Machine Translation on Sentiment Classification. In *Proceedings of Association for Machine Translation in the Americas, AMTA*, pages 78–88, Orlando, Florida.

Alberto Poncelas, Wichaya Pidchamook, Chao-Hong Liu, James Hadley, and Andy Way. 2020b. Multiple Segmentations of Thai Sentences for Neural Machine Translation. In *Proceedings of The 1st Joint Spoken Language Technologies for Under-resourced languages and Collaboration and Computing for Under-Resourced Languages Workshop, SLTU-CCURL*, pages 240–244, Marseille, France.

Alberto Poncelas, Maja Popovic, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019b. Combining SMT and NMT back-translated data for efficient NMT. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 922–931, Varna, Bulgaria.

Maja Popović and Alberto Poncelas. 2020. Extracting correctly aligned segments from unclean parallel data using character n-gram matching. In *Konferenca Jezikovne tehnologije in digitalna humanistika*, pages 74–80, Ljubljana, Slovenia.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and Effective Subword Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Seattle, USA.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 86–96, Berlin, Germany.

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

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

Kaveh Taghipour, Nasim Afhami, Shahram Khadivi, and Saeed Shiry. 2010. A discriminative approach to filter out noisy sentence pairs from bilingual corpora. In *Proceedings of 5th International Symposium on Telecommunications (IST 2010)*, pages 537–541, Tehran, Iran.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, (LREC)*, pages 2214–2218, Istanbul, Turkey.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 484–491, Rochester, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, USA.

Oscar Wilde, Alfred Douglas, trans, and Hendrik Bulthuis, trans. 1891 (1894, 1910). *Salomé (Salome)*. Wikisource, Project Gutenberg.