Prompsit's Submission to the IWSLT 2018 Low Resource Machine Translation Task

Víctor M. Sánchez-Cartagena

Prompsit Language Engineering
Av. Universitat s/n. Edifici Quorum III
E-03202 Elx, Spain
vmsanchez@prompsit.com

Abstract

This paper presents Prompsit Language Engineering's submission to the IWSLT 2018 Low Resource Machine Translation task. Our submission is based on cross-lingual learning: a multilingual neural machine translation system was created with the sole purpose of improving translation quality on the Basque-to-English language pair. The multilingual system was trained on a combination of in-domain data, pseudo in-domain data obtained via cross-entropy data selection and backtranslated data. We morphologically segmented Basque text with a novel approach that only requires a dictionary such as those used by spell checkers and proved that this segmentation approach outperforms the widespread byte pair encoding strategy for this task.

1. Introduction

This paper presents Prompsit Language Engineering's submission to the IWSLT 2018 Low Resource Machine Translation task. The objective of this task is building an MT system for translating TED talks from Basque to English from a very limited amount of in-domain Basque–English parallel data. We relied on cross-lingual learning via a multilingual approach [1] to neural machine translation (NMT), extraction and cleaning of pseudo in-domain parallel text from out-of-domain data, and backtranslation of Spanish text into Basque for building our submission.

Moreover, we applied morphological segmentation to the Basque text. We took advantage of an existing spell checking dictionary and its inflection paradigms and used an automatic morphology inference model to decide between ambiguous segmentations. We proved that this method, that requires shallower linguistic information[1] than other segmentation approaches based on full morphological analysis and disambiguation [2, 3], outperforms the widespread byte pair encoding (BPE) segmentation strategy [4] in terms of translation quality for Basque-to-English NMT.

[1]Neither part of speech/morphological information in the dictionary nor a part of speech tagger/parser are needed. In principle, this approach could be applied to any language for which a Hunspell-based (`http://hunspell.github.io/`) spell checker exists.

Table 1: *Size of in-domain data. Processed segments are those that remain after removing talks included in the development and test sets.*

Language pair	# raw segments	# processed segments
eu-en	5 687	5 687
eu-es	6 742	5 610
eu-fr	7 021	5 878
es-en	280 947	279 737
fr-en	290 961	289 722

The remainder of the paper explains the steps followed to build the submitted NMT system. Next section explains how the in-domain and out-of-domain parallel corpora were processed and filtered, while Section 3 focuses on describing and assessing the impact of the morphological segmentation approach followed. Section 4 describes the NMT architecture and training process. Section 5 depicts the process followed to obtain the data set used to train our submission. Finally, the most relevant related approaches are reviewed in Section 6 and the paper ends with some concluding remarks.

2. Data acquisition and cleaning

Our submission was trained on a combination of in-domain and out-of-domain data. The only special cleaning applied to the in-domain training data provided by the organization is the removal of talks that are also included in the test/development sets. Table 1 shows the number of segments in the in-domain data for each language pair before and after removing such talks.

Following the shared task instructions,[2] we built the out-of-domain data collection by downloading all the corpora available from the Opus [5] and WMT [6] websites.[3] We also included the Basque–Spanish parallel data from Open

[2]`https://sites.google.com/site/iwsltevaluation2018/TED-tasks`

[3]If the same corpus was available from both websites (e.g. Europarl), we downloaded it from WMT. If the same corpus was available from different WMT editions, we downloaded it from the most recent one. We skipped some corpora from Opus which were too noisy, like EUBookshop.

Table 2: *Size of out-of-domain data before and after applying shallow cleaning.*

Language pair	# raw segments	# clean segments
eu-en	1.81M	928K
eu-es	1.64M	1.41M
eu-fr	711K	375K

Table 3: *Size of out-of-domain data before and after applying aggressive cleaning.*

Language pair	# raw segments	# clean segments
es-en	163M	65M
fr-en	164M	77M

Data Euskadi Repository published by the task organizers.

We followed two different strategies for out-of-domain parallel data cleaning. For language pairs with limited data availability, namely those including Basque, we followed a conservative shallow cleaning strategy since removing correct segments can be harmful for the quality of the final system. For the remaining language pairs, since only a subset of the data is finally used (see Section 5), we applied a more aggressive cleaning strategy.

The shallow cleaning consisted in deduplication and removal of parallel segments that meet any of the following conditions: they contain a low proportion of alphabetic characters, their source-language (SL) and target-language (TL) side are very similar (there is a low edit distance between them), they are too long or too short (shorter than 3 tokens or longer than 100), or they are written in another language (language is detected by means of `cld2`[4] and segments are only discarded when the language detection is *reliable* according to the `cld2` algorithm). Table 2 shows the size of the out-of-domain data for each language pair containing Basque before and after applying shallow cleaning.[5]

The aggressive cleaning consisted in two steps. Firstly, parallel segments were deduplicated and a more aggressive superset of the rules used in the shallow cleaning (implemented in the translation memory cleaning tool `Bicleaner`[6]) was applied. These rules are addressed at detecting evident flaws such as encoding errors, very different lengths in parallel segments, etc. Secondly, misaligned segments were detected and removed by means of an automatic classifier, described in [7]. The classifier is also part of the `Bicleaner` tool. Pre-trained models for the classifier were obtained from the Paracrawl project.[7] Table 3 shows the size of the out-of-domain data for each language pair before and after applying the aggressive cleaning.

[4] `https://github.com/CLD2Owners/cld2`

[5] Shallow cleaning was not applied to the Basque–Spanish parallel data from Open Data Euskadi Repository.

[6] `https://github.com/bitextor/bicleaner`

[7] `https://github.com/bitextor/bitextor-data/tree/master/bicleaner`

3. Morphological segmentation for Basque

Word segmentation based on linguistically-informed strategies such as morphological analysis [2] or simpler alternatives based on lists of relevant prefixes and suffixes [8] have shown to be able to outperform the popular BPE approach [4] for some agglutinative and highly inflected languages. In this section, we present the pseudo-morphological segmentation approach based on inflection paradigms we applied to Basque text in our submission and prove that it outperforms BPE.

3.1. Pseudo-morphological segmentation based on inflection paradigms

Inflection paradigms are commonly used in dictionaries (morphological dictionaries used in rule-based machine translation, spell checkers, etc.) in order to group regularities in the inflection of a set of words.[8] A paradigm is usually defined as a collection of suffixes and, optionally, their corresponding part-of-speech/morphological information; e.g., the paradigm assigned to many common English verbs indicates that by adding the suffix *-ing* to the stem, the gerund is obtained; by adding the suffix *-ed*, the past is obtained; etc. While morphological dictionaries from rule-based machine translation systems contain morphological information, spell checkers usually lack this information.

In languages with a high inflection degree, such as Basque, a surface form can be built by sequentially appending suffixes from different paradigms to a stem. For instance, the word *etxekoak* can be generated from the entry $etxe + PAR240$ if paradigm $PAR240$ contains the suffixes $-ko + PAR243$, $-z$, $-rekin$, etc. and paradigm $PAR243$ contains the suffix $-ak$.

As suffixes contained in inflection paradigms are usually based on linguistic knowledge, one can take advantage of inflection paradigms for splitting words for training NMT systems. In this way, words can be split in atomic units of meaning or *morphs*. For instance, in previous example, *etxekoak* (the plural form of "domestic") would be split into *etxe* ("house"), *-ko* (adjectivation) and *-ak* (plural mark).

In order to split a corpus using inflection paradigms, there are two types of words for which an additional strategy needs to be devised:

- Homograph words: those that can be generated by multiple combinations of stem and suffix(es).
- Unknown words: those that are not present in the morphological dictionary/spell checking dictionary.

In order to decide the best segmentation for these words, we took advantage of semi-supervised morphology learning methods. In particular, we used Morfessor [9]. Morfessor is

[8] Paradigms ease dictionary management by reducing the quantity of information that needs to be stored, and by simplifying revision and validation because of the explicit encoding of regularities in the dictionary.

a family of methods for automatic learning of morphology based on the minimum description length principle [10]: the words in a corpus are split in morphs in such a way that the size of the morph vocabulary and the length in tokens of the corpus are minimized. We used a semi-supervised variant of Morfessor in which the segmentation model can be estimated from a plain corpus and a set of already segmented words [11].

Our pseudo-morphological segmentation strategy comprises the following steps:

1. Segment words encoded in the morphological dictionary/spell checker which have only a candidate segmentation according to the inflection paradigms.

2. Train a Morfessor segmentation model in an semi-supervised way [11] from the Basque corpus we want to segment and the words segmented in the previous step.

3. Segment homograph words by choosing the segmentation with the highest likelihood according to the previous model.

4. Segment unknown words by choosing the segmentation with highest likelihood according to the model among those that can be generated by using solely suffixes from the inflection paradigms in the morphological dictionary/spell checker.

This approach hence allows us to segment a corpus in atomic units of meaning using a spell checker as the only linguistic resource. Unlike other approaches to NMT training corpus word segmentation based on linguistic information [2, 3], this approach does not require neither a full morphological analyzer with part-of-speech/morphological tags nor a part-of-speech tagger/parser for disambiguating between the different analyses of each word. Part-of-speech/morphological information (e.g. the fact the suffix *-ed* represents the past tense of a verb) is not used during the process and disambiguation is carried out by the Morfessor model which, in turn, controls the growth of vocabulary size.

In our submission, we used the Basque spell checker *Xuxen v5.1* as dictionary.[9] Moreover, following [8], we applied BPE splitting with a model learned on the concatenation of all training corpora after performing the pseudo-morphological segmentation. Note that applying BPE to further split the word pieces obtained after pseudo-morphological segmentation helps the system to translate proper nouns and compounds in Basque.

3.2. Evaluation

We evaluated the pseudo-morphological segmentation approach we employed in our submission and compared it with two baselines: a greedy alternative in which the segmentation with the most frequent stem is chosen for unknown and homograph words, and plain BPE splitting. In all cases, BPE was applied to all the languages of the training corpus (65 000 operations) and the model was learned from their concatenation after carrying out pseudo-morphological segmentation (except for the plain BPE system, for which morphological segmentation was not carried out).

[9] `https://xuxen.eus`

Table 4: *Results of the evaluation of the pseudo-morphological segmentation approach proposed, a greedy alternative, and plain BPE.*

Segmentation strategy	BLEU	TER
BPE	12.75	83.68
Paradigms/Greedy+BPE	13.28	87.80
Pseudo-morph+BPE	13.59	79.73

We trained multilingual NMT systems as described in Section 4 on parallel corpora segmented following the three strategies. The three multilingual NMT systems were trained on the in-domain data and included the language pairs Spanish–English, French–English, Basque–English, Basque–French and Basque–Spanish.

The evaluation was carried out only on the Basque-to-English direction. The values of the translation evaluation metrics BLEU [12] and TER [13] computed on the development set are reported in Table 4. We can observe that our pseudo-morphological segmentation approach (`Pseudo-morph+BPE`) outperforms both plain BPE segmentation and segmentation based on paradigms with a greedy strategy for homograph and unknown words.

Table 5 shows several examples of words segmented by the three alternatives evaluated. Furthermore, Table 6 depicts three Basque sentences from the development set, how they were segmented by the three alternatives evaluated and their translation with the NMT systems built. Note that, unlike the words in Table 5, the SL sentences in Table 6 were split with BPE after applying the splitting strategies based on inflection paradigms, as described previously in this section. In the first example, the Basque word *konpartimentutan* is formed by the stem *konpartimentu*, which means "compartment", plus the inessive suffix *-tan*). The segmentation strategies based on inflection paradigms are able to correctly detach the inessive suffix from the word, while the pure BPE approach fails to do it. As a consequence, the MT system built using the latter approach is not able to produce an adequate translation by taking advantage of the sentences in the training corpus that contain words starting with *konparti-*. Similarly, in the second example, the segmentation strategies based on inflection paradigms are able to segment *estudioa* into the stem *estudio* (that means "studio apartment") and the suffix *-a* (singular article). The pure BPE approach segments it into *estudi-* and *-oa*. Since *estudi-* is the stem of the verb "to study" in Spanish, the multilingual system wrongly generates that verb in the translation into English. Finally, in the third example, the greedy approach based on paradigms wrongly segments *Asia* into *as* and *-ia*, which prevents the NMT system from

Table 5: *Examples of Basque words segmented by the three approaches evaluated. The segmentation that best splits the word in atomic units of meaning is shown in bold.*

Word	BPE	Paradigms/Greedy	Pseudo-morph	meaning
adierazitako	adierazitako	adieraz itako	**adierazi tako**	"expressed", built from *adierazi* ("to express") plus *-tako* (suffix used in relative clauses)
izendatu	**izendatu**	izenda tu	**izendatu**	"nominate", atomic unit
ebaluaketa	ebalu aketa	ebaluaket a	**ebaluaketa**	"evaluation", atomic unit
birgaitzeko	bir gaitzeko	**birgaitze ko**	**birgaitze ko**	"rehabilitation" (*birgaitze*) plus genitive suffix (*-ko*)

producing the word *Asia* in English.

4. Training strategy

Our submission is based on cross-lingual learning. We aimed at improving the translation performance on the Basque-to-English language pair by means of the addition of training data from other language pairs. The different language pairs were combined by means of a multilingual NMT approach [1]. A TL marker was prepended to each SL segment. See Section 5 for more details about language pairs included and how the data for each of them was obtained.

Our submitted NMT system follows the Transformer architecture [14]. In particular, we used the implementation in the Marian NMT toolkit [15]. We generally used the hyperparameters of the *Transformer base model* [14], with the exception of *warmup_steps*, which was set to 16 000 instead of 4 000. This parameter was increased because our minibatch size was significantly smaller than that used in the original paper [14]. We limited segment length to 100 tokens and let the Marian toolkit set the batch size to fit 8 000 MiB of GPU memory. For a vocabulary size of around 70 000 words, the number of TL words in a minibatch was around 3 000, while [14] report 25 000 TL words per minibatch. A checkpoint was saved every 5 000 updates.

We used only the publicly released Basque–English *IWSLT18.TED.dev2018* corpus as a development set.[10] Training ended when perplexity on the development set did not improve in 10 consecutive checkpoints. We selected the checkpoint that obtained the highest BLEU score on the development set.

Concerning corpora preprocessing, text was tokenized with the `aggressive` strategy[11] implemented by the OpenNMT tokenizer [16]. Words were split in sub-word units as described in Section 3. The Morfessor model was trained on the concatenation of the Basque section of the training data for all language pairs that contained Basque. The BPE model (65 000 operations), which shared by all SLs and TLs, was learned from the concatenation of the morphologically segmented Basque data and the unsegmented data for the remaining languages and it was used to split these corpora. Text was lowercased prior to training and the resulting English translations were recased[12] with a recasing model estimated from the concatenation of the English side of the training corpora.

[10] It could be interesting to study whether using development data from other language pairs has a significant impact in translation quality for Basque–English.

[11] The only multi-character tokens allowed are sequences of strictly alphabetical characters.

5. Training data

This section describes the training data from which our submission was built and the experiments carried out to select it.

5.1. Language pairs

According to the experiments carried out by [1], including new language pairs that share either the SL or the TL with the language pair of interest helps to increase the translation quality for that language pair. Henceforth, our multilingual system contains only language pairs with Basque as SL or English as TL. Moreover, we included only language pairs for which the training set is published as part of this year's data. Hence, our multilingual system contains data from the Spanish–English, French–English, Basque–English, Basque–French and Basque–Spanish language pairs. Preliminary experiments showed no important gains when adding data from the German–English and Turkish–English language pairs to the training collection. Conducting more exhaustive experiments has been left as future work.

5.2. Cross-entropy data selection and oversampling

As shown in Table 3, there is a huge amount of out-of-domain parallel data available for the Spanish–English and French–English language pairs. If it was just concatenated to the in-domain data, the system would be biased towards the out-of-domain data. In order to avoid that issue, we selected only a subset of the out-of-domain data which is similar to the in-domain one (from now on, *pseudo in-domain data*) via cross-entropy difference [17].

The process was carried out as follows. Firstly, we sorted the out-of-domain data (after cleaning it as described in Section 2) by monolingual cross-entropy difference on the English side. The in-domain language model was estimated

[12] The Moses recaser was used: `http://www.statmt.org/moses/?n=Moses.SupportTools#ntoc10`.

Table 6: Result of applying each of the three segmentation strategies evaluated in Section 3 to a three sentences extracted from the development set. The translation of each sentence with a multilingual NMT system trained only on the in-domain data is also depicted. The character → at the end of a token implies that it is a sub-word unit originally attached to the token that follows it. Words whose segmentation has a visible impact on the translation are shown in bold.

#	segmentation strategy	sentence
1	source – BPE	burmu→ ina ez dago **kon**→ **parti**→ **men**→ **tutan** ban→ atuta .
	source – Paradigms/Greedy+BPE	bur→ mu→ in→ a ez dago **kon**→ **parti**→ **mentu**→ **tan** bana→ tuta .
	source – Pseudo-morph+BPE	bur→ mu→ in→ a ez dago **kon**→ **parti**→ **mentu**→ **tan** bana→ tuta .
	translation – BPE	There's no brain at all based on **bias**.
	translation – Paradigms/Greedy+BPE	You don't have a brain that's broken up into **blocks**.
	translation – Pseudo-morph+BPE	There's no **boundary** in the brain.
	reference	The brain isn't divided into **compartments**.
2	source – BPE	beraz urte batez **estudi**→ **oa** ix→ tea erabaki nuen .
	source-Paradigms/Greedy+BPE	bera→ z urte bat→ ez **estudio**→ **a** ix→ te→ a erabaki nu→ en .
	source – Pseudo-morph+BPE	beraz urte bat→ ez **estudio**→ **a** ix→ te→ a erabaki nuen .
	translation – BPE	So I decided to **study** for a year.
	translation – Paradigms/Greedy+BPE	So one year I decided to give it a try.
	translation – Pseudo-morph+BPE	So I decided to stay silent for a year.
	reference	So I decided to close it down for one year.
3	source – BPE	beraz **asia** aukeratu nuen .
	source – Paradigms/Greedy+BPE	bera→ z **as**→ **ia** aukera→ tu nu→ en .
	source – Pseudo-morph+BPE	beraz **asia** aukeratu nuen .
	translation – BPE	So I chose **Asia**.
	translation – Paradigms/Greedy+BPE	So I decided to give it a try.
	translation – Pseudo-morph+BPE	So I chose **Asia**.
	reference	So **Asia** it was.

from the English side of the parallel in-domain Spanish–English training corpus, while the out-of-domain one was obtained from a random sample with the same number of segments from the English side of all the available Spanish–English parallel data. The same language models were used for computing monolingual cross-entropy difference for both the Spanish–English and French–English language pairs. As other authors did previously [18], we split English corpora with BPE prior to training the language models and scoring the out-of-domain parallel segments.

Secondly, we carried out a set of experiments in order to decide which is the most appropriate amount of pseudo in-domain data for Spanish–English and French–English. In these experiments, we used all the available data for Basque–English, Basque–Spanish and Basque–French, and varying amounts of pseudo in-domain data, which was concatenated to the real in-domain data, for Spanish–English and French–English. In addition, we also studied the effect of oversampling the Basque–English data (concatenation of in-domain and out-of-domain) to match the size of the Spanish–English and French–English data.

Table 7 depicts the size of the pseudo in-domain parallel data[13] and the size of the Basque–English data included in the training set for the different configurations evaluated, together with the values of the evaluation metrics BLEU [12] and TER [13] computed on the development set. The original size of the Basque–English data is 933 356 segments. Those rows with values higher than 0.9M imply that the data the Basque–English has been oversampled. In other words, it has been included as many times as necessary for reaching the size depicted in the table. Systems were trained following the set-up described in Section 4. For the same data configurations, Table 8 shows automatic evaluation metrics computed after finetuning the systems on the in-domain data.[14] Results show no important gains when increasing the out-of-domain data size from 2M to 5M and confirm the importance of oversampling, in line with the results reported in [1]. Finetuning on in-domain data did not bring any positive impact. One possible reason could be the scarce amount of in-domain Basque–English data available (see Table 1). We chose the configuration with the highest BLEU score on the development set (depicted in bold in Table 7) for our submission.

5.3. Backtranslation

Backtranslation, that is, the translation of additional TL monolingual data into the SL with an MT system in order to obtain additional training material, is a widespread method to enhance the quality of NMT systems [19].

[13] For a given size N, the N parallel segments with the lowest cross-entropy score are selected from the out of domain data.

[14] When finetuning, the initial learning rate was set to the value employed in the last update of the main training process.

Table 7: *Results of the experiments carried out in order to determine the best size for pseudo in-domain data and for Basque–English data (with oversampling). Unlike the experiments depicted in Table 8, these experiments did not include finetuning on the in-domain data at the end of the training process. The configuration highlighted in bold is the one used in our submission.*

Pseudo in-domain size	eu-en size	BLEU	TER
2M	0.9M	21.05	68.15
2M	**2M**	**21.72**	**68.65**
5M	0.9M	19.47	71.86
5M	3M	20.46	70.17
5M	5M	21.10	68.95

Table 8: *Results of the experiments carried out in order to determine the best size for pseudo in-domain data and for Basque–English data (with oversampling). Unlike the experiments depicted in Table 7, these experiments included finetuning on the in-domain data at the end of the training process.*

Pseudo in-domain size	eu-en size	BLEU	TER
2M	0.9M	21.15	69.11
2M	2M	21.68	68.46
5M	0.9M	19.96	71.20
5M	3M	20.88	71.46
5M	5M	21.68	69.38

In our submission, we did not directly translate monolingual English data into Basque. Since there is high-quality Basque–Spanish parallel data not available for Basque–English (Open Data Euskadi Repository) we opted for translating the Spanish side of Spanish–English parallel data into Basque in order to build additional Basque–English training material. A similar approach has been successfully applied for enhancing phrase-based statistical machine translation systems [20].

In order to carry out the backtranslation, we trained an NMT system on all the available Spanish–Basque data with the set-up described in Section 4. Words were segmented as described in Section 3. That system was used to backtranslate the Spanish side of the in-domain Spanish–English training data and the top 5M segments[15] from the pseudo in-domain Spanish–English corpus.

We evaluated the impact of adding backtranslated data to the best dataset from the previous section (2M pseudo in-domain parallel segments, oversampling and no finetuning). We built NMT systems after adding the full backtranslated data (both the in-domain and the pseudo-in-domain data; row labeled as *5.2M*), and after adding the in-domain and only 2M pseudo-in-domain backtranslated segments (row labeled as *2.2M*). Results of the evaluation on the development set of the NMT systems trained with these data are depicted in Table 9. They show that using the whole backtranslated data has a strong positive impact on the quality of the resulting MT system. Hence, we used the 5.2M backtranslated segments in our submission.

[15]We could not backtranslate a larger amount of data because of time restrictions.

Table 9: *Results of the experiments carried out in order to determine the best size for backtranslated data. The configuration highlighted in bold is the one used in our submission.*

Size of backtranslated data	BLEU	TER
0	21.72	68.65
2.2M	22.51	67.54
5.2M	**23.45**	**66.94**

5.4. Final submission

Our final submission was trained on the best data collection from previous section. We experimented with finetuning and checkpoint ensembling [21, Sec. 3.2], but translation quality did not improve. Hence, we submitted just the result of translating the test set with the intermediate model that achieved the highest BLEU score on the development set.

6. Related approaches

Our submission is built with the help of morphological segmentation, cross-entropy data selection and cross-lingual learning via multilingual NMT. This section reviews the most relevant approaches in these three fields.

Morphological segmentation has been successfully applied to build a winning system [2] for the English–Finnish language pair in the WMT 2016 news translation shared task [22]. Simpler alternatives based on lists of prefixes/suffixes have also been reported to bring improvements in translation quality [8]. Morphological segmentation has already been applied to NMT for Basque [23]. However, unlike our approach, their strategy segments homograph words in a greedy way (longest stem). Besides morphological segmentation, there are other ways linguistic resources can be used to segment words for NMT training. For instance, TL words can be transformed into a sequence of stem and morphological inflection tags in order to achieve better morphological generalization when translating into highly inflected languages [3].

Cross-entropy data selection [17] has become a popular approach for leveraging out-of-domain data when building MT systems. This strategy has been used for collecting training data for phrase-based statistical machine translation systems [24] and NMT systems [18] in shared translation tasks such as WMT [6] and IWSLT [25].

In multilingual NMT [1], a single NMT model is used to translate between different language pairs. Some authors proposed multilingual NMT strategies in which the underly-

ing network architecture does not need to be modified [1, 26]. That property allowed us to perform multilingual MT with a Transformer [14] model despite the fact that the multilingual NMT approach we followed [1] was originally addressed to the encoder-decoder with attention architecture [27]. On the contrary, other authors [28] proposed modifying the network architecture to use an independent encoder and decoder for each language.

7. Concluding remarks

This paper presented Prompsit Language Engineering's submission to the IWSLT 2018 Low Resource MT track. We presented a novel method for morphological segmentation based solely on a dictionary with inflection paradigms such as those used by spell checkers and proved that it outperforms the widespread BPE segmentation method. Our submission relies on cross-lingual learning via multilingual NMT. Basque training data was segmented with the novel method. The NMT system follows the Transformer architecture. We experimented with varying amounts of pseudo in-domain data obtained via cross-entropy data selection and with varying amounts of backtranslated data and submitted the combination that maximized translation quality on the development set.

Our submission could be further improved with independent ensembles [21, Sec. 3.2]. The inclusion of additional language pairs has not been exhaustively evaluated and the quality of the final system might be improved by adding some more language pairs. The quality of the final system could also improve with the addition of more backtranslated data.

8. Acknowledgements

We would like to thank Prof. Mikel L. Forcada for the advice on Basque segmentation. Work supported by project IADAATPA, action number 2016-EU-IA-0132, funded under the Automated Translation CEF Telecom instrument managed by INEA at the European Commission.

9. References

[1] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, *et al.*, "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association of Computational Linguistics*, vol. 5, no. 1, pp. 339–351, 2017.

[2] V. M. Sánchez-Cartagena and A. Toral, "Abu-matran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, vol. 2, 2016, pp. 362–370.

[3] A. Tamchyna, M. Weller-Di Marco, and A. Fraser, "Modeling target-side inflection in neural machine translation," in *Proceedings of the Second Conference on Machine Translation*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017, pp. 32–42. [Online]. Available: http://www.aclweb.org/anthology/W17-4704

[4] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 1715–1725.

[5] J. Tiedemann, "Parallel Data, Tools and Interfaces in OPUS," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[6] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, *et al.*, "Findings of the 2017 conference on machine translation (WMT17)," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 169–214.

[7] V. M. Sánchez-Cartagena, M. Bañón, S. Ortiz-Rojas, and G. Ramírez-Sánchez, "Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task," in *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*. Brussels, Belgium: Association for Computational Linguistics, October 2018.

[8] M. Huck, S. Riess, and A. Fraser, "Target-side word segmentation strategies for neural machine translation," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 56–67.

[9] S. Virpioja, P. Smit, S.-A. Grönroos, and M. Kurimo, "Morfessor 2.0: Python implementation and extensions for morfessor baseline, D4 Julkaistu kehittmis- tai tutkimusraportti tai -selvitys," 2013. [Online]. Available: http://urn.fi/URN:ISBN:978-952-60-5501-5

[10] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465 – 471, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0005109878900055

[11] O. Kohonen, S. Virpioja, and K. Lagus, "Semi-supervised learning of concatenative morphology," in *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, ser. SIGMORPHON '10. Stroudsburg,

PA, USA: Association for Computational Linguistics, 2010, pp. 78–86. [Online]. Available: http://dl.acm.org/citation.cfm?id=1870478.1870488

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[13] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[15] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 116–121. [Online]. Available: http://www.aclweb.org/anthology/P18-4020

[16] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "Opennmt: Open-source toolkit for neural machine translation," *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, 2017.

[17] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 355–362. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145474

[18] M. Junczys-Dowmunt and A. Birch, "The University of Edinburgh's systems submission to the MT task at IWSLT," in *Proceedings of IWSLT 2016*, 2016.

[19] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2016, pp. 86–96.

[20] M. Huck and H. Ney, "Pivot lightly-supervised training for statistical machine translation," in *Proc. 10th Conf. of the Association for Machine Translation in the Americas*, 2012, pp. 50–57.

[21] R. Sennrich, A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A. V. M. Barone, and P. Williams, "The University of Edinburgh's Neural MT Systems for WMT17," in *Proceedings of the Second Conference on Machine Translation*, 2017, pp. 389–399.

[22] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, *et al.*, "Findings of the 2016 conference on machine translation," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, vol. 2, 2016, pp. 131–198.

[23] T. Etchegoyhen, E. Martínez Garcia, A. Azpeitia, G. Labaka, I. Alegria, I. Cortes Etxabe, A. Jauregi Carrera, I. Ellakuria Santos, M. Martin, and E. Calonge, "Neural Machine Translation of Basque," in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, Alacant, Spain, 2018, pp. 139–148.

[24] R. Rubino, A. Toral, V. M. Sánchez-Cartagena, J. Ferrández-Tordera, S. O. Rojas, G. Ramírez-Sánchez, F. Sánchez-Martínez, and A. Way, "Abu-MaTran at WMT 2014 translation task: Two-step data selection and RBMT-style synthetic rules," in *Proceedings of the ninth workshop on statistical machine translation*, 2014, pp. 171–177.

[25] C. Mauro, F. Marcello, B. Luisa, N. Jan, S. Sebastian, S. Katsuitho, Y. Koichiro, and F. Christian, "Overview of the IWSLT 2017 Evaluation Campaign," in *International Workshop on Spoken Language Translation*, 2017, pp. 2–14.

[26] T.-L. Ha, J. Niehues, and A. Waibel, "Toward multilingual neural machine translation with universal encoder and decoder," in *Proceedings of 2016 International Workshop on Spoken Language Translation*, 2016.

[27] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[28] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural machine translation with a shared attention mechanism," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 866–875.