A multi-source approach for Breton–French hybrid machine translation

Víctor M. Sánchez-Cartagena, Mikel L. Forcada, Felipe Sánchez-Martínez
Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
E-03690 Sant Vicent del Raspeig (Spain)
{vmsanchez,mlf,fsanchez}@dlsi.ua.es

Abstract

Corpus-based approaches to machine translation (MT) have difficulties when the amount of parallel corpora to use for training is scarce, especially if the languages involved in the translation are highly inflected. This problem can be addressed from different perspectives, including data augmentation, transfer learning, and the use of additional resources, such as those used in rule-based MT (RBMT). This paper focuses on the hybridisation of RBMT and neural MT (NMT) for the Breton–French under-resourced language pair in an attempt to study to what extent the RBMT resources help improve the translation quality of the NMT system. We combine both translation approaches in a multi-source NMT architecture and find out that, even though the RBMT system has a low performance according to automatic evaluation metrics, using it leads to improved translation quality.

1 Introduction

Corpus-based approaches to machine translation (MT), such as neural MT (NMT), struggle when the size of the available parallel corpora for a given language pair is scarce (Koehn and Knowles, 2017). Even though the problem can be partially mitigated with accurate hyper-parameter tuning (Sennrich and Zhang, 2019), taking advantage of additional resources can help to further improve the quality of the system.

Monolingual texts in both languages can be leveraged with the help of back-translation (Sennrich et al., 2016a; Hoang et al., 2018) to generate synthetic parallel corpora. It is also possible to use only monolingual corpora and follow an unsupervised NMT approach (Artetxe et al., 2018). Parallel corpora from related language pairs can also be leveraged thanks to multilingual NMT (Johnson et al., 2017) and other forms of transfer learning (Kocmi and Bojar, 2018).

In addition to the use of corpora, linguistic resources can also be used to improve NMT. If morphological analysers or syntactic parsers are available, they can be used to build a richer representation of the words being translated (Sennrich and Haddow, 2016; Nadejde et al., 2017). Even full rule-based MT (RBMT) systems can be combined with NMT in order to build hybrid systems (Huang et al., 2020).

In this work, we focus on an under-resourced language pair: Breton–French, and study mechanisms to build a hybrid system by combining NMT with the Breton–French system built with the Apertium RBMT platform (Forcada et al., 2011).

We aim at producing sentences that combine knowledge extracted from the parallel corpus and from the RBMT system. Hence, we go beyond approaches that simply choose the best system (either RBMT or NMT) for each input sentence (see below). We use multi-source NMT and formalise the problem of combining both sources of knowledge as an automatic post-editing (Chatterjee et al., 2018) problem. In this way, we are able to explore different ways of generating the RBMT output, using different resources, to study which resources are more useful for the hybrid approach.

The rest of the paper is organised as follows. The remainder of this section lists previous works related to the hybridisation of RBMT and corpus-based systems, including approaches for integrating external bilingual segments into NMT. Section 2 then explains the resources available for Breton–

Martins, Moniz, Fumega, Martins, Batista, Coheur, Parra, Trancoso, Turchi, Bisazza, Moorkens, Guerberof, Nurminen, Marg, Forcada (eds.)
Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, p. 61–70
Lisboa, Portugal, November 2020.

French and the challenges of translating between Breton and French. Section 3 describes the hybrid architecture chosen. Section 4 presents the experiments carried out and discusses the results obtained. The paper ends with some concluding remarks.

Hybrid systems combining rule-based and corpus-based approaches. The creation of hybrid systems combining RBMT and statistical MT (SMT) has been explored by many authors. The most relevant approach for this work (Tyers, 2009) enlarged the training corpus of an SMT system with 116,500 *sentence pairs* made up of all possible inflected Breton forms and their inflected French translations as present in an earlier version of the Apertium Breton–French system we are using. Schwenk et al. (2009) followed a similar approach for other language pairs. More sophisticated approaches (Eisele et al., 2008; Enache et al., 2012; Sánchez-Cartagena et al., 2016) involve modifying the SMT architecture.

Concerning the combination of RBMT and NMT, a relevant line of research involves choosing the best output (either RBMT or NMT) for each source sentence. For instance, Huang et al. (2020) propose training an automatic classifier for this task and use some features to help predict how difficult is the source sentence for each system: for instance, the degree of morphological and syntactic ambiguity is useful to estimate how difficult is the sentence for the RBMT system, while the token frequency on the training corpus can help to assess how difficult it is for the NMT system. Similarly, Singh et al. (2019) use confidence scores computed for each system to choose the best alternative for each source sentence. Torregrosa et al. (2019) experimented with the integration of RBMT bilingual dictionaries and syntactic parsers into NMT without success.

Finally, the multi-source architecture studied in this paper has been preliminary explored by Sánchez-Cartagena et al. (2019). The main differences with this work are: i) they did not study the impact of the different components of the RBMT system; and ii) they did not perform a hyper-parameter search, which could explain the poor performance of their transformer systems. In addition, we conduct an automatic analysis of the errors produced by our hybrid approach.

Integration of bilingual segments into NMT. The integration of bilingual segments, which could be produced by an RBMT system, into an NMT system has received some attention recently. One of the first approaches (Arthur et al., 2016), which can only be applied to single-token bilingual segments, used the attention weights of a recurrent attentional encoder–decoder (Bahdanau et al., 2015) model to decide the target language (TL) word translation probabilities that needed to be boosted in the final softmax layer. Tang et al. (2016) and Wang et al. (2017) relied on a phrase memory for NMT that could contain multiple-token bilingual segments. They modelled decoding as a mixture of two processes: generating a word with the standard NMT model, or introducing a phrase from the phrase memory. Zhang et al. (2017) formalised the strategy of Tang et al. (2016) as a *posterior regularization* approach (Ganchev et al., 2010). Feng et al. (2018) designed a phrase attention mechanism that could be used either without additional supervision or with an external bilingual lexicon. Another related line of research modifies the beam search algorithm to meet some terminological constraints (Chatterjee et al., 2017; Post and Vilar, 2018).

2 Breton–French machine translation

The Breton language (*Brezhoneg* in Breton) is a Celtic language of the Brittonic group that is spoken in the west of Brittany (*Breizh Izel* or “Lower Brittany”) in France, and the main language with which it has contact is French, the only official language; in fact, Breton, spoken by about 200,000 people, has virtually no legal recognition in France.

Resources for Breton: Programs like Firefox, Google applications and some Microsoft programs have been localized and there is a 70,000-page Breton Wikipedia. There is little software dedicated to Breton; most of it free/open-source, such as the Apertium MT system and the LanguageTool spelling and grammar checker. This software and services such as the Freelang online dictionary[1] are based on linguistic resources such as morphological analyzers, monolingual and bilingual dictionaries. As for bilingual text corpora, today OPUS[2] contains about 400,000 sentence pairs, most of them very specialized, in the field of computer science.

The Apertium Breton–French system: The Apertium platform[3] contains an MT system designed to allow French-speaking readers to access written Breton content (*gisting*).[4] This MT system

[1] https://www.freelang.com/enligne/breton.php

[2] http://opus.nlpl.eu

[3] http://www.apertium.org

[4] Developers deliberately chose not develop French–Breton MT, deeming it too risky in terms of the socio-linguistic situation, as users would assume the machine-translated Breton to be

(Tyers, 2010), the only one in the world for Breton, was released in May 2009 as the result of the joint efforts of the *Ofis ar Brezhoneg*,[5] the Spanish company Prompsit Language Engineering, and the Universitat d'Alacant and is based on the Apertium platform (Forcada et al., 2011). Dictionary development started with the free dictionaries for Breton in Lexilogos.[6] Development of the Apertium Breton–French MT system slowly continues. The quality of the French generated is not suitable for publishing, but may be used to get a rough idea of the meaning of a Breton text.

Automatic inference of translation rules for Breton–French: There have been attempts to improve the Apertium Breton–French system in an unsupervised way. In particular, Sánchez-Cartagena et al. (2015) proposed an algorithm for the automatic inference of shallow-transfer rules from small parallel corpora and existing RBMT dictionaries.The result of applying the algorithm to the Apertium Breton–French system using just the parallel data prepared by Tyers (2009) was a set of rules whose quality, as measured by automatic MT evaluation metrics, was close to the existing hand-crafted ones.

3 System architecture

We propose combining the explicit linguistic knowledge encoded in the Breton–French Apertium system with the implicit knowledge encoded in a parallel corpus by means of multi-source NMT (Zoph and Knight, 2016). Given a source-language (SL) sentence to be translated, our proposed architecture proceeds as follows (see Figure 1): First, the SL sentence is translated with the RBMT system; then the original SL sentence and its RBMT translation are passed as inputs to the multi-source NMT system, which produces the final translation. At training time, the SL side of the parallel sentences in the training corpus is translated with Apertium to obtain a "trilingual" parallel corpus. As it is common practice, the multi-source system works on byte-pair-encoding (BPE) sub-word units (Sennrich et al., 2016b) obtained from both inputs and the output together.

With this architecture, we expect the NMT system to learn to translate from the SL text with help from the RBMT output. It could also be seen the other way round: the NMT system postedits the RBMT output with the help of the SL sentence. In fact, this architecture has been successfully applied for automatic post-editing (Junczys-Dowmunt and Grundkiewicz, 2018).

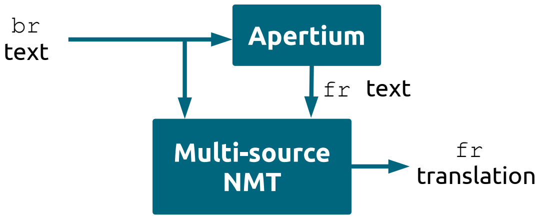

Figure 1: Multi-source NMT approach followed to integrate the linguistic knowledge encoded in the Apertium Breton–French RBMT system.

The Apertium architecture as well as the multi-source NMT architecture used in our experiments are described in the remainder of this section.

3.1 Apertium rule-based machine translation

Apertium is a free/open-source RBMT system that follows a shallow-transfer architecture. What follows is brief description of its modules; for a complete description of the system we refer the reader to the work by Forcada et al. (2011).

- A *morphological analyser* segments the text in surface forms (*words*, or, where detected, multi-word lexical units) and delivers, for each one, one or more *lexical forms* consisting of lemma, lexical category and morphological inflection information.
- A *part-of-speech tagger*, which combines a constraint grammar (Karlsson et al., 1995) with a first-order hidden Markov model (Cutting et al., 1992), selects the most likely lexical form corresponding to an ambiguous surface form.
- A *lexical transfer* module which reads each SL lexical form and delivers the corresponding TL lexical form by looking it up in a bilingual dictionary.
- A shallow *structural transfer* module that performs syntactic operations on the sequence of lexical forms to improve the grammaticality of the output.[7]
- A *morphological generator* which delivers a TL surface form for each TL lexical form, by suitably inflecting it.

good and use it improperly as if it were correct (Jakez, 2009 personal communication).

[5] Now *Ofis Publik ar Brezhoneg*

[6] https://www.lexilogos.com/breton_dictionnaire.htm

[7] This shallow model does not rely on a full parse tree of the whole sentence and, therefore, RBMT systems that perform full syntactic analysis are more effective than Apertium when dealing, for instance, with long-range reorderings.

- A *post-generator* which performs inter-word orthographic operations: contractions, elisions marked by apostrophes, etc.[8]

3.2 Multi-source neural machine translation

We experimented with the transformer (Vaswani et al., 2017) and the recurrent attentional encoder–decoder (Bahdanau et al., 2015, hereinafter recurrent) NMT architectures. In both cases, we followed the multi-source architectures implemented in the Marian toolkit (Junczys-Dowmunt et al., 2018), which are described next.

Our recurrent NMT systems follow the same architecture as Nematus (Sennrich et al., 2017b), namely a bidirectional gated recurrent unit (GRU) encoder, a conditional GRU decoder with attention (Miceli Barone et al., 2017, Sec. 4.2) and a deep output that combines the context vector, the recurrent hidden state and the embedding of the previous symbol. The multi-source recurrent NMT system contains two encoders (one for each input) which do not share parameters. The modifications in the decoder that allow it to accommodate the two encoders are the following:

- The initial state of the decoder is obtained after concatenating the averaged encoder states of the two input sequences.
- The conditional GRU (cGRU) unit with attention in the decoder is replaced by a doubly-attentive cGRU cell (Calixto et al., 2017) featuring two independent attention mechanisms.
- The context vector used in the deep output is replaced by the concatenation of the context vectors of the two inputs.

For further details, the reader is referred to Junczys-Dowmunt and Grundkiewicz (2017).

Our transformer models follow the architecture proposed by Vaswani et al. (2017). A transformer model contains an encoder and a decoder. The encoder is made of stacked layers, each containing a self-attention unit and a feed-forward unit. The decoder is also made of stacked layers, each containing a self-attention unit, an encoder–decoder attention unit and feed-forward unit. The multi-source transformer systems contain two encoders and two encoder–decoder attention units in each decoder layer. This transformer multi-source architecture was also used in the winning submission to the 2018 WMT automatic post-editing shared task (Chatterjee et al., 2018). For further details, the reader is referred to Junczys-Dowmunt and Grundkiewicz (2018).

Corpus	# sent.	# br tokens	# fr tokens
train	139,489	1,096,311	1,116,100
dev	2,000	25,291	24,835
test	3,000	37,054	36,346

Table 1: Number of parallel sentences and tokens in Breton and French for the corpora used for train/dev/test corpora.

4 Experiments and results

For the experiments we used the following corpora available at OPUS:[9] Tatoeba, GNOME, OfisPublik, KDE4, wikimedia, Ubuntu and OpenSubtitles. For development and testing we used the same portions of the OfisPublik corpus used by Sánchez-Cartagena et al. (2015), the rest of corpora, after de-duplication, were used for training. Table 1 reports the amount of parallel sentences and tokens in each language for the training, development, and test corpora.

Concerning Apertium, we used the Breton–French data available at `https://github.com/apertium/apertium-br-fr`. In addition to the shallow-transfer rules included in these linguistic data, we also experimented with shallow-transfer rules automatically inferred from the portion of the OfisPublik corpus included in the training corpus using the algorithm by Sánchez-Cartagena et al. (2015).

In order to determine the appropriate amount of BPE operations and hyper-parameter values to be used for the two models we proceed as follows: First we tried with 5,000, 10,000, 20,000, and 30,000 BPE operations with a baseline system not using any Apertium data. When doing so the rest of hyper-parameters were set to the values recommended by Sennrich et al. (2017a) for the recurrent model and by Vaswani et al. (2017) for the base transformer model, respectively, except for the model size which was set to 512. Training stopped after 5 validations without any perplexity improvement on the development corpus; validations were performed every 1,000 mini-batches; each minibatch contained 8,000 tokens. The best results were obtained with 20,000 BPE operations for the recurrent model and 5,000 for the transformer. We then performed a grid search to find the appropriate hyper-parameters for each model. The hyper-parameters tried for the recurrent model are:

- Embedding sizes in $\{512, 256, 128\}$. For each embedding size the hidden size was set to twice the size of the embeddings.

[8] In French: *à* + *lequel* → *auquel*; *de* + *hôtels* → *d'hôtels*, etc.

[9] `http://opus.nlpl.eu`

- Encoder and decoder cell depths in $\{1, 2, 4, 8\}$. We used the same value for both so as not to explore the Cartesian product. Cell depth is defined as the number of GRU transitions in the deep transition architecture proposed by Miceli Barone et al. (2017, Sec. 4.2).

The hyper-parameters tried for the transformer model are:

- Attention heads in $\{2, 4, 8\}$.
- Model size in $\{512, 256, 128\}$.
- Encoder and decoder layers in $\{2, 4, 6\}$. As before, we used the same value for both to avoid exploring the Cartesian product.

The best results for the recurrent model were obtained with an embedding size of 512 and encoder and decoder cell depths of 2. For the transformer, the best results were obtained with 4 attention heads, model size of 512 and 4 encoder and decoder layers. These hyper-parameters are the ones used for the rest of experiments reported.

Table 2 provides the BLEU and chrF2++ scores for the reference systems and for the different ways of exploiting the linguistic resources in Apertium, as explained next. For the reference NMT systems and the different multi-source NMT configurations we have tried, the table reports the mean and standard deviation of the scores obtained after three different training executions.

An explanation of the different reference systems follows:

- Baseline NMT system (*base NMT*) trained solely on the training corpus (see Table 1).
- Baseline NMT system trained on a concatenation of the training corpus and the entries in the Breton–French bilingual dictionary of Apertium (*base+dic NMT*). Tyers (2009) explains how all the inflected bilingual entries can be obtained from the Apertium dictionaries; some of them may have more than one translation equivalent while others may be multiword entries. The amount of bilingual entries obtained from the current version is 125,829, of which 57 have more than one translation equivalent and 2,228 are multiword entries.
- Apertium with hand-crafted rules (*RBMT man. rules*): the full RBMT system. The linguistic resources used by this system are: morphological analyser for Breton, morphological generator for French, part-of-speech tagger of Breton, Breton–French bilingual dictionary of lemmas and shallow structural transfer rules.
- Apertium with automatically-inferred rules (*RBMT auto rules*). Same as above but using the shallow structural transfer rules automatically inferred by Sánchez-Cartagena et al. (2015), instead of using hand-crafted rules.
- Apertium with no structural transfer rules (*RBMT no rules*). Same as above but using no structural transfer rules. After morphological analysis and part-of-speech tagging the lexical forms in Breton are translated into lexical forms in French one by one, without applying any structural transfer to make the output more grammatical, except for very simple one-word rules that ensure that the morphological features sent to the French generator for each separate word are valid.

As regards the different ways of exploiting the linguistic resources in Apertium, we generated the additional input translation provided to the multi-source NMT system with the same RBMT configurations used as reference systems (see above) as well as a word-for-word translation obtained using exactly the same bilingual dictionary we used for the *base+dic NMT* reference system. As this dictionary contains multi-word lexical units, we translated word for word in a left-to-right, longest-match fashion so that the bilingual entry covering the longest sequence of tokens is selected when there is more than one possibility. When the bilingual dictionary contained more than one translation per source word, they were all included in the output separated by a special token. This happened to 495 source words in the training corpus.

The results in Table 2 show that the use of Apertium resources improves translation quality according to both BLEU and chrF2++. The best improvement, about 1.3 BLEU points, is obtained when the additional input to the multi-source NMT system is obtained without structural transfer rules (*RBMT no rules*). However, if we pay closer attention to the performance of the reference system *RBMT no rules* on its own, the scores it obtains are worse than those obtained with hand-crafted rules (*RBMT man. rules*) and automatically inferred rules (*RBMT auto rules*). This results suggest that Apertium may be helping the NMT system to perform a better lexical selection, since the improvement in the grammaticality of the Apertium output provided by the shallow-transfer rules has no effect on the quality of the final translation. In any case, the use of a morphological analyser and part-of-speech tagger for Breton has a positive effect on the translation quality of the multi-source NMT system; compare the

BLEU	Recurrent	Transformer
reference systems		
base NMT	21.25 ± 0.12	18.45 ± 0.08
base+dic NMT	21.26 ± 0.24	18.50 ± 0.15
RBMT man. rules	12.45	
RBMT auto rules	12.16	
RBMT no rules	8.78	
multi-source		
RBMT man. rules	21.36 ± 0.46	19.16 ± 0.02
RBMT auto rules	22.24 ± 0.46	19.48 ± 0.18
RBMT no rules	**22.59 ± 0.06**	19.70 ± 0.15
word-for-word	21.73 ± 0.22	18.24 ± 0.13

chrF2++	Recurrent	Transformer
reference systems		
base NMT	38.38 ± 0.13	36.94 ± 0.03
base+dic NMT	38.68 ± 0.13	37.25 ± 0.09
RBMT man. rules	35.16	
RBMT auto rules	33.86	
RBMT no rules	30.91	
multi-source		
RBMT man. rules	39.58 ± 0.27	38.80 ± 0.08
RBMT auto rules	40.12 ± 0.34	39.03 ± 0.15
RBMT no rules	**40.49 ± 0.10**	39.19 ± 0.17
word-for-word	39.20 ± 0.10	37.17 ± 0.17

Table 2: BLEU and chrF2++ evaluation scores for different reference systems and for the different multi-source NMT configurations we have tried. RBMT stands for the Apertium rule-based MT used.

performance of *RBMT no rules* with the *word-for-word* translation which uses a bilingual dictionary of surface forms. Finally, the addition of the bilingual dictionary to the training corpus seems to have no effect on translation quality.

In order to get a deeper insight about the effect of the different hybridisation strategies, we carried out an automatic error analysis following the strategy of Toral and Sánchez-Cartagena (2017). We used Hjerson (Popović, 2011), [10] which classifies errors into five word-level categories: inflection errors, reordering errors, missing words, extra words and incorrect lexical choices. As it is difficult to automatically distinguish between the latter three categories (Popović and Ney, 2011), we grouped them into a unique category named *lexical errors*. Hjerson works on the surface form and lemma of the words in the reference translations and MT outputs. The lemmas used were obtained with the StandfordNLP lemmatiser (Qi et al., 2018).

We computed the relative difference in the number of Hjerson errors in the test set between the multi-source NMT systems and the *base NMT* system;[11] a positive value means that the multi-source system made more errors than the *base NMT* system. Table 3 shows, for the recurrent and transformer architectures, the relative difference computed for each error category and for the total number of errors. As each training was repeated 3 times, the table reports the average and standard deviation of the relative difference for the 9 possible combinations between training runs. In order to contextualise the relative differences, Table 4 reports the average and standard deviation of the total number of errors of each type in the baseline system.

For the recurrent architecture, the addition of expanded dictionaries to the bilingual training corpus does not significantly alter the number of errors. One possible explanation could be that the potential gains of introducing more lexical knowledge in the system are neutralised by the presence of single-word sentences in the training corpus, that could harm the fluency of the generated sentences.

Multi-source NMT systems, on the contrary, tend to make fewer lexical errors than the *base NMT* system. This happens for three out the four multi-source systems, where the system with hand-crafted rules is the only one in which the reduction in lexical errors is not statistically significant. Neither automatically inferred nor hand-crafted transfer rules cause a statistically significant impact in the amount of inflection errors, and both of them make reordering errors increase. The multi-source system without transfer rules is the best performing system according to automatic evaluation metrics because it is the one that brings the largest reduction in lexical errors, which constitute the most frequent error category (see Table 4). It is worth noting that the bilingual dictionary in Apertium contains a single translation for each SL lexical form, hence its lexical selection capabilities are poor. Overall, it seems that the multi-source system is able to make a better use of the translations from the bilingual dictionary when they are sequentially placed in the additional input rather than when they have been processed by transfer rules.

Concerning the transformer architecture, some differences in the way the different error categories change can be observed. The transformer seems to be more robust to the addition of dictionaries to the training corpus: adding them leads to a statistically significant reduction in lexical errors. Moreover, the transformer multi-source systems make more

[10] `https://github.com/cidermole/hjerson`

[11] Computed as $\frac{\#\text{errors\_multi\_source} - \#\text{errors\_base}}{\#\text{errors\_base}}$.

Recurrent	inflection	reordering	lexical	total
reference systems				
base+dic NMT	-0.019 ± 0.024	-0.022 ± 0.022	0.012 ± 0.024	0.007 ± 0.020
multi-source				
RBMT man. rules	0.006 ± 0.020	0.031 ± 0.017	-0.017 ± 0.032	-0.011 ± 0.028
RBMT auto rules	-0.015 ± 0.028	0.039 ± 0.025	-0.049 ± 0.024	-0.040 ± 0.021
RBMT no rules	0.008 ± 0.016	0.045 ± 0.018	**-0.066 ± 0.031**	**-0.052 ± 0.026**
word-for-ford	-0.005 ± 0.021	0.005 ± 0.022	-0.030 ± 0.027	-0.025 ± 0.023

Transformer	inflection	reordering	lexical	total
reference systems				
base+dic NMT	-0.010 ± 0.015	-0.012 ± 0.017	-0.009 ± 0.004	-0.010 ± 0.003
multi-source				
RBMT man. rules	0.048 ± 0.018	**0.112 ± 0.017**	-0.014 ± 0.006	0.001 ± 0.005
RBMT auto rules	0.048 ± 0.018	0.093 ± 0.019	-0.024 ± 0.004	-0.010 ± 0.003
RBMT no rules	**0.060 ± 0.016**	0.092 ± 0.032	-0.023 ± 0.003	-0.008 ± 0.004
word-for-ford	0.007 ± 0.021	-0.003 ± 0.018	-0.005 ± 0.004	-0.004 ± 0.003

Table 3: For each NMT architecture, average and standard deviation of the relative changes in the amount of errors for each error category (inflection, reordering, lexical and total). Increases in the amount of error whose confidence interval does not intersect with zero are shown in red, decreases whose confidence interval does not intersect with zero are shown in green. For each error type, the largest relative change is shown in **bold**.

	Recurrent	Transformer
inflection	1971 ± 27	1869 ± 27
reordering	2969 ± 44	2910 ± 42
lexical	30641 ± 726	27599 ± 84

Table 4: For each architecture, absolute number of errors for each type detected by the Hjerson tool on the translation of the test set with the baseline NMT system.

inflection and reordering errors than the recurrent ones. Nevertheless, the lexical errors behave in a similar way in both multi-source architectures: the configuration that leads to the largest reduction in the number of lexical errors is the RBMT system with no transfer rules.

Table 5 shows how the different systems evaluated translate a few sentences from the test set. In the first example, the baseline system is not able to correctly translate the Breton words *e-barzh* and *e-maez*, whose meaning is correctly captured by the Apertium dictionaries. The multi-source systems are able to produce the right translations (*entrées* and *sorties*, respectively *entrances* and *exits* in English) or at least related words, while the *base+dic NMT* repeats *entrées*. In the second example, whose sentence structure is more complex, the baseline system fails to produce a translation that conveys the meaning of the fragment of the reference *On leur a donné le nom de satellites galiléens, en hommage à Galilée*, which roughly means *They were given the name of Galilean satellites, in homage to Galileo*. Only two hybrid systems were able to generate a translation that captures that meaning of the fragment: the multi-source systems without transfer rules and with automatically inferred rules.

5 Concluding remarks

This paper focused on the hybridisation of RBMT and NMT for the Breton–French under-resourced language pair. The aim of the paper is to study to what extent the resources from the Apertium RBMT system help the NMT system to improve its output. We combined both translation approaches in a multi-source NMT architecture and explore the use of different resources in the Apertium Breton–French system to generate the RBMT translation to be used as an additional input.

Despite the low performance of the RBMT system, the hybrid system is able to outperform a pure NMT baseline. The best translation performance is achieved with a hybrid system whose RBMT subsystem contains no transfer rules at all but takes advantage of the Breton morphological analyser and part-of-speech tagger, the French generator and post-generator and the bilingual dictionary.

The fact that the use of no transfer rules provides the best results while the RBMT system using no transfer rules, when evaluated in isolation, performs worse than the rest of RBMT configurations may seem contradictory. However, the automatic error analysis revealed that the hybrid systems using

#	system	sentence
1	source	Staliañ panelloù divyezhek evit **mont e-barzh ha mont e-maez** ar gumun.
	baseline	mise en place d'une signalétique bilingue sur **le site internet** de la commune.
	RBMT no rules	Installer panneaux bilingues pour **aller à l'intérieur et aller hors** de le commune.
	RBMT auto rules	Installer panneaux bilingues pour **aller à l'intérieur et aller hors** de la commune.
	RBMT man. rules	Installer des panneaux bilingues pour **aller à l'intérieur et aller hors** de la commune.
	base+dic NMT	Installation de panneaux bilingues **à l'entrée et de l'entrée** de la commune.
	ms. word-for-word	Mise en place des panneaux bilingues aux **entrées et sorties** de la commune.
	ms. RBMT no rules	Mise en place de panneaux bilingues pour **entrer et sortie** de la commune.
	ms. RBMT auto rules	Il s'agit pour l'installation de panneaux bilingues aux **entrées et sorties** de la commune.
	ms. RBMT man. rules	Installation de panneaux bilingues **d'entrée et de sortie** d'agglomération.
	reference	Mise en place de panneaux bilingues aux **entrées et sorties** de la commune.
2	source	Adplanedennoù galilean a vez graet anezho e koun Galileo Galilei, ar steredoniour italian a zizoloas anezho e 1610 gant ul lunedenn hepken.
	baseline	Les satellites galiléens Galilei, l'astronome italien redécouvre en 1610 avec un œil nu.
	RBMT no rules	Satellites galilean a être faire d'eux dans mémoire Galileo Galilei, le astronome italienne a découvrir d'eux dans 1610 avec un lunette seulement.
	RBMT auto rules	Satellites galilean qui les faire des en mémoire Galileo Galilei, le astronome italien qui découvrir des à 1610 par une lunette seulement.
	RBMT man. rules	Satellites galilean Il est fait d'eux dans mémoire Galileo Galilei, l'astronome italien découvrit d'eux dans 1610 avec une lunette seulement.
	base+dic NMT	Les satellites galiléens sont des satellites galiléens, dont l'astronome italien découvre en 1610 à un œil nu.
	ms. word-for-word	Les satellites galiléens de Galilée, l'astronome italienne traversent en 1610 par une lunette uniquement.
	ms. RBMT no rules	**Satellites galiléens sont évoqués dans la mémoire Galileo Galilei, l'astronome italienne** vous découvrira en 1610 avec une lunette unique.
	ms. RBMT auto rules	**De plus, les satellites galiléens forment la mémoire Galileo** qui les découvre en 1610 par une lunette unique.
	ms. RBMT man. rules	Les satellites galiléens, l'astronome italien découvrit en 1610 par une lunette seulement.
	reference	**On leur a donné le nom de satellites galiléens, en hommage à Galilée (astronome Italien)** qui les découvrit en 1610 avec une simple lunette.

Table 5: Translations into French of different Breton sentences extracted from the test set and produced by the different hybrid strategies evaluated (recurrent architecture; *ms.* stands for *multi-source*). The most remarkable differences are highlighted.

no transfer rules make fewer lexical errors, which account for most of the errors produced by the systems, but more reordering and inflection errors.

Since transfer rules seem not to be needed in our multi-source approach to succeed and morphological analysers, morphological generators and small bilingual dictionaries are available for many under-resourced language pairs, we hope that the hybrid approach presented in this paper opens the door to the development of more accurate hybrid systems in under-resource scenarios.

Acknowledgments: Work funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 825299, project Global Under-Resourced Media Translation (GoURMET). We thank NVIDIA Corporation the donation of one of the GPUs used for training.

References

Artetxe, M., G. Labaka, E. Agirre, and K. Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*, Vancouver, Canada, May.

Arthur, P., G. Neubig, and S. Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, November.

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May.

Calixto, I., Q. Liu, and N. Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada, July.

Chatterjee, R., M. Negri, M. Turchi, M. Federico, L. Specia, and F. Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark, September.

Chatterjee, R., M. Negri, R. Rubino, and M. Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 723–738, Belgium, Brussels, October.

Cutting, D., J. Kupiec, J. Pedersen, and P. Sibun. 1992. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 133–140, Trento, Italy.

Eisele, A., C. Federmann, H. Saint-Amand, M. Jellinghaus, T. Herrmann, and Y. Chen. 2008. Using Moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 179–182, Columbus, Ohio, USA, June.

Enache, R., C. España Bonet, A. Ranta, and L. Màrquez Villodre. 2012. A hybrid system for patent translation. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 269–276, Trento, Italy, May.

Feng, J., L. Kong, P.-S. Huang, C. Wang, D.Huang, J. Mao, K. Qiao, and D. Zhou. 2018. Neural phrase-to-phrase machine translation. *CoRR*, abs/1811.02172.

Forcada, M.L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J.A. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F.M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Ganchev, K., J. Graça, J. Gillenwater, and B. Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Hoang, V.C.D., P. Koehn, G. Haffari, and T. Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July.

Huang, J.-X., K.-S. Lee, and Y.-K. Kim. 2020. Hybrid translation with classification: Revisiting rule-based and neural machine translation. *Electronics*, 9(2).

Johnson, M., M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Viégas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Junczys-Dowmunt, M. and R. Grundkiewicz. 2017. An exploration of neural sequence-to-sequence architectures for automatic post-editing. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 120–129, Taipei, Taiwan, November.

Junczys-Dowmunt, M. and R. Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 822–826, Belgium, Brussels, October.

Junczys-Dowmunt, M., 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. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July.

Karlsson, F., A. Voutilainen, J. Heikkilä, and A. Anttila. 1995. Constraint grammar: A language-independent system for parsing unrestricted text. mouton de gruyter.

Kocmi, T. and O. Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium, October.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada, August.

Miceli Barone, A.V., J. Helcl, R. Sennrich, B. Haddow, and A. Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107, Copenhagen, Denmark, September.

Nadejde, M., S. Reddy, R. Sennrich, T. Dwojak, M. Junczys-Dowmunt, P. Koehn, and A. Birch. 2017. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*, pages 68–79, Copenhagen, Denmark, September.

Popović, M. and H. Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Popović, M. 2011. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–67.

Post, M. and D. Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, USA, June.

Qi, P., T. Dozat, Y. Zhang, and C.D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. 2015. A generalised alignment template formalism and its application to the inference of shallow-transfer machine translation rules from scarce bilingual corpora. *Computer Speech & Language*, 32(1):46 – 90.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. 2016. Integrating rules and dictionaries from shallow-transfer machine translation into phrase-based statistical machine translation. *Journal of Artificial Intelligence Research*, 55(1):17–61.

Sánchez-Cartagena, V. M., J. A. Pérez-Ortiz, and F. Sánchez-Martínez. 2019. The Universitat d'alacant submissions to the English-to-Kazakh news translation task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 356–363, Florence, Italy, August.

Schwenk, H., S. Abdul-Rauf, L. Barrault, and J. Senellart. 2009. SMT and SPE machine translation systems for WMT'09. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 130–134, Athens, Greece, March.

Sennrich, R. and B. Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany, August.

Sennrich, R. and B. Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July.

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

Sennrich, R., B. Haddow, and A. 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, August.

Sennrich, R., A. Birch, A. Currey, U. Germann, B. Haddow, K. Heafield, A.V. Miceli Barone, and P. Williams. 2017a. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark, September.

Sennrich, R., O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A.V. Miceli Barone, J. Mokry, and M. Nadejde. 2017b. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April.

Singh, M., R. Kumar, and I. Chana. 2019. Improving neural machine translation using rule-based machine translation. In *Proceedings of the 7th International Conference on Smart Computing Communications*, pages 1–5, Miri, Malaysia, June.

Tang, Y., F. Meng, Z. Lu, H. Li, and P.L.H. Yu. 2016. Neural machine translation with external phrase memory. *CoRR*, abs/1606.01792.

Toral, A. and V. M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Valencia, Spain, April.

Torregrosa, D., N. Pasricha, M. Masoud, B. R. Chakravarthi, J. Alonso, N. Casas, and M. Arcan. 2019. Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 125–133, Dublin, Ireland, August.

Tyers, F.M. 2009. Rule-based augmentation of training data in Breton–French statistical machine translation. In *Proceedings of the 13th Annual Conference of the European Association of Machine Translation*, pages 213–218, Barcelona, Spain, May.

Tyers, F.M. 2010. Rule-based Breton to French machine translation. In *Proceedings of the 14th Annual Conference of the European Association of Machine Translation*, pages 174–181, Saint-Raphaël, France, May.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Wang, X., Z. Tu, D. Xiong, and M. Zhang. 2017. Translating phrases in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Copenhagen, Denmark, September.

Zhang, J., Y. Liu, H. Luan, J. Xu, and M. Sun. 2017. Prior knowledge integration for neural machine translation using posterior regularization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1514–1523, Vancouver, Canada, July.

Zoph, B. and K. Knight. 2016. Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, CA, USA, June.