Understanding the effects of word-level linguistic annotations in under-resourced neural machine translation

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez

Dep. de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
E-03690 Sant Vicent del Raspeig (Spain)
{vmsanchez,japerez,fsanchez}@dlsi.ua.es

Abstract

This paper studies the effects of word-level linguistic annotations in under-resourced neural machine translation, for which there is incomplete evidence in the literature. The study covers eight language pairs, different training corpus sizes, two architectures and three types of annotation: dummy tags (with no linguistic information at all), part-of-speech tags, and morpho-syntactic description tags, which consist of part of speech and morphological features. These linguistic annotations are *interleaved* in the input or output streams as a single tag placed before each word. In order to measure the performance under each scenario, we use automatic evaluation metrics and perform automatic error classification. Our experiments show that, in general, source-language annotations are helpful and morpho-syntactic descriptions outperform part of speech for some language pairs. On the contrary, when words are annotated in the target language, part-of-speech tags systematically outperform morpho-syntactic description tags in terms of automatic evaluation metrics, even though the use of morpho-syntactic description tags improves the grammaticality of the output. We provide a detailed analysis of the reasons behind this result.

1 Introduction

Training neural machine translation (NMT) systems for under-resourced language pairs, for which the amount of parallel corpora is orders of magnitude smaller than those available for prevailing language pairs, may be challenging. Recently, Sennrich and Zhang (2019) have shown that even under these circumstances, NMT surpasses classical approaches such as phrase-based statistical MT (Koehn, 2010). In these under-resourced scenarios, the use of relevant linguistic word-level annotations has proved to improve translation performance (Sennrich and Haddow, 2016; Nadejde et al., 2017).

Linguistic annotations can be used to label source-language (SL) or target-language (TL) words. The former lead to more accurate representations of the SL sentence (Sennrich and Haddow, 2016) and usually require changes in the encoder but not in the training loss (Chen et al., 2017; Li et al., 2018). The latter involve producing probability distributions for both TL words and TL linguistic annotations, which can be seen as a form of multi-task learning. Multi-task learning architectures explored in the literature include: independent decoders for words and linguistic annotations (Zhou et al., 2017; Wu et al., 2018; Gū et al., 2018; Wang et al., 2018; Yang et al., 2019); independent output layers in the same decoder (García-Martínez et al., 2016; Grönroos et al., 2017; Feng et al., 2019); and even sharing the same network for both tasks (Nadejde et al., 2017; Tamchyna et al., 2017; Wagner, 2017) and alternatively produce linguistic annotations and words. The latter approach is usually referred to as *interleaving*.

We can also classify the approaches according to the type of linguistic annotations used: part-of-speech tags (Feng et al., 2019; Yang et al., 2019); morpho-syntactic description tags, which comprise part of speech and morphological inflection information (García-Martínez et al., 2016; Tamchyna et al., 2017); and syntactic structure information (Nadejde et al., 2017; Chen et al., 2017; Zhou et al., 2017; Wu et al., 2018; Gū et al., 2018; Wang et al., 2018). Using morpho-syntactic description tags as TL annotations

allows us to train the network to produce TL lemmas instead of surface forms (words as they appear in running texts). This strategy can reduce data sparseness but requires the use of an external morphological generator as a post-processing step (Tamchyna et al., 2017).

Despite the body of work published about this topic, no strategy has clearly emerged as the most appropriate method for integrating linguistic annotations into NMT. The literature mainly contains incomplete evidence. For instance, Yang et al. (2019) conclude that TL part-of-speech annotations boost translation quality with an *ad-hoc* architecture, but Wagner (2017) claims that TL morpho-syntactic description tags degrade translation quality when they are interleaved: it is not clear whether the difference between both results is caused by the type of linguistic annotations or by the approach followed to integrate them. There are also contradictory results, such as those reported by Tamchyna et al. (2017), who claim that TL annotations are only useful when they are combined with lemmatisation, and Nadejde et al. (2017), who report positive results without lemmatisation. In addition, the influence of factors such as the size of the available training parallel corpus and the language typology have not been properly evaluated.

In this paper, we aim at clarifying how linguistic annotations help NMT by carrying out systematic experiments with eight language pairs. We focus on an under-resourced scenario where linguistic annotations are likely to provide information that cannot be inferred from scarce training data. We analyse multiple factors, namely, language typology, side which is annotated with linguistic information (SL, TL or both), architecture of the NMT system, training corpus size, and type of information encoded in the tags. For the latter factor, we focus only on part-of-speech tags and morpho-syntactic description tags, since other type of annotations, such as CCG supertags (Steedman, 2000), are unlikely to be available for under-resourced languages. We train systems on linguistically annotated surface forms via interleaving, which does not require modifications to the neural network and allows us to easily compare different types of linguistic annotations and NMT architectures. In addition, an automatic error classification allows us to qualitatively compare them. A qualitative analysis has only been previously performed by Nadejde et al. (2017), but it covered only two language pairs with English as TL, and a single type of annotation.

The rest of the paper is organised as follows. Next section overviews the process of interleaving linguistic annotations in the training data. Section 3 then describes the experimental settings whereas Section 4 reports and discusses the results obtained. Section 5 presents automatic error classification results for all the systems evaluated, while Section 6 studies the reasons behind the poor performance of systems with TL morpho-syntactic description tags. The paper ends with some concluding remarks.

2   Interleaving in neural machine translation

The interleaving approach for integrating linguistic annotations into NMT (Nadejde et al., 2017) annotates each word with a single tag which is *interleaved* in the sentence before the word, i.e. introduced in the sentence as if it were another word. In our experiments, tags can represent either the part of speech (POS) of the word or its morpho-syntactic description (MSD). As corpora are pre-processed with BPE (Sennrich et al., 2016b), the tag is introduced just once, before the first sub-word unit. To study if the effect of using tags is related to the fact that input and output sequences get longer and word boundaries are explicitly defined, and not to the information provided by tags, we also tried with a dummy tag (DUM) conveying no linguistic information at all, and used the same dummy tag for every word (Wagner, 2017). Interleaved TL tags are removed from the final translation generated by the system before computing the automatic evaluation metrics.

The example below shows the result of interleaving MSD tags in the English sentence *It has happened before*. The sentence contains a pronoun (PRON) followed by an auxiliary verb (AUX), a main verb (VERB), an adverb (ADV) and a punctuation mark (PUNCT). The analysis of the pronoun tells us that it is personal, nominative, neuter, singular, and 3rd-person. The symbol @@ acts as a sub-word unit separator.

```
PRON_Case=Nom|Gender=Neut|Number=Sing|Person=3|PronType=Prs it
AUX_Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin has
VERB_Tense=Past|VerbForm=Part happen@@ ed ADV__ before PUNCT__ .
```

3 Experimental settings

We conducted experiments for the translation of English text into four languages, and vice-versa. These languages —Czech (cs), German (de), Spanish (es) and Turkish (tr)— belong to different language families and differ at the syntactic and morphological levels. German, Czech and Spanish are Indo-European languages: they are, respectively, Germanic, Slavic and Romance. Both German and Czech have declension and SVO sentence structure, except for the subordinate sentences in German, which are SOV. Spanish has no declension and its sentence structure is SVO. Turkish is an agglutinative Turkic language with declension and SOV sentence structure. The morphological differences between these languages are reflected in the sparsity of the MSD tags: the number of unique tags in the interleaved training corpora ranges from a few hundreds for English and Spanish to a few thousands for Czech, German and Turkish.[1]

We simulated an under-resourced scenario by downsampling available parallel corpora for the selected language pairs. Downsampling has important advantages over using truly under-resourced language pairs: (i) we can choose languages from different families and evaluate them on standard, high-quality test sets; (ii) we can confirm whether conclusions hold for richer-resource scenarios by training on larger datasets for the same language pairs; and (iii) linguistic annotations can be obtained with the same state-of-the-art morphological analyser, minimising the potential distortions introduced by differences in the morphological analyser technology and in performance between the languages. The POS and MSD tags were obtained by means of the StandfordNLP tagger (Qi et al., 2018). In any case, the approaches described in this paper could be applied to truly under-resourced language pairs as transfer learning allows to obtain morphological analysers even from scarce morphologically annotated data (Kondratyuk, 2019).

Corpora. The training, development and test sets used belong to the news domain. For training, we used texts from the News Commentary v14 corpus,[2] except for Turkish, for which we used texts from the SETimes corpus (Tyers and Alperen, 2010). For development and testing we used evaluation sets from the WMT 2019 Conference on Machine Translation, each of which contains around 3,000 parallel sentences.[3] To see if the conclusions drawn on the under-resourced settings hold in a richer-resourced scenario, we trained English–German systems (in both directions) on the concatenation of the parallel data made available for the WMT 2017 shared task on news translation[4] plus the synthetic parallel data obtained through back-translation released by Sennrich et al. (2016a).

All corpora were tokenised and truecased with the Moses scripts[5] and parallel sentences longer than 100 words in either side were discarded. Table 1 provides information about the training corpora after their pre-processing. We trained translation models on these corpora and on random sub-sets of them containing 50k parallel sentences (except for the WMT training data). The token counts depicted in Table 1 for the under-resourced scenario are similar to those listed in the OPUS collection (Tiedemann, 2012) for under-resourced language pairs such as English–Kurdish or English–Igbo; token counts for the 50k subsets match other pairs with even smaller resources available in OPUS, such as English–Kazakh.

Translation models. We tested the performance of the recurrent-neural-network encoder-decoder with attention (hereafter, *recurrent*; Bahdanau et al., 2015) and the Transformer (Vaswani et al., 2017) architectures when the different types of tags introduced in Section 2 are interleaved in the SL input sequence, in the TL output sequence, and in both of them. For each architecture, we also trained a baseline using no tags at all. To keep the experiments to a manageable size, the systems that included tags in both languages were not trained on the nine possible combinations of tag types (three in the SL and three in the TL). Instead, they were trained only on SL MSD and TL POS tags, which were those with the best general performance when used in isolation in the SL and in the TL, respectively. For the same reason, we only explored the SL MSD/TL POS tag combination for systems trained on large-scale WMT data.

---

[1]The exact figures are as follows: English: 208; Czech: 1,904; German: 1,077; Spanish: 365; Turkish: 1,859.

[2]http://data.statmt.org/news-commentary/v14/

[3]For Turkish, German and Czech we used newstest2017 and newstest2018 for development and testing, respectively. For Spanish, we used newstest2012 for development and newstest2013 for testing.

[4]http://data.statmt.org/wmt17/translation-task/preprocessed

[5]https://github.com/moses-smt/mosesdecoder/tree/master/scripts

language pair	tokens	sentences	language pair	tokens	sentences
under-resourced			**WMT**		
English–Czech	6.0M – 5.4M	240k	English–German	212M – 200M	9.4M
English–German	8.4M – 8.4M	329k	German–English	232M – 240M	10.0M
English–Spanish	9.5M – 10.6M	367k			
English–Turkish	5.2M – 4.7M	208k			

Table 1: Number of sentences and tokens in the parallel corpora used for training. The column labelled as *tokens* depicts the number of tokens for each language of the pair, in the order defined by the content of the column *language pair*. Figures for the WMT data differ between both directions because of the use of backtranslated corpora.

In order to determine the appropriate values for training hyper-parameters, a grid search over the number of BPE operations and the neural network sizes was carried out. The optimum hyper-parameter values for each language pair, training corpus size and architecture were obtained after training the baseline systems. These hyper-parameters were also used with the systems integrating linguistic annotations. Appendix A provides a detailed description of the training process.

Error classification. We followed the automatic error analysis strategy by Toral and Sánchez-Cartagena (2017), who used the tool Hjerson (Popović, 2011) to classify word errors into five categories: inflection, reordering, 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 were obtained again with the StandfordNLP tagger.

4 Results and discussion

Table 2 shows the BLEU (Papineni et al., 2002) scores obtained by the different systems. A score in bold means that the system outperforms the baseline (labelled as *None*) by a statistically significant margin. A bullet (•) next to the score of a system with interleaved POS or MSD tags means that it outperforms the system with DUM tags in the same language side (SL or TL) by a statistically significant margin.[6] A dagger (†) next to the score of a system with POS or MSD tags means that it outperforms the system with the opposite tag (either MSD or POS) in the same language side by a statistically significant margin. Statistical significance was assessed with paired bootstrap resampling (Koehn, 2004) ($p = 0.05$; 1 000 iterations).

As the four languages paired with English are morphologically richer than English, we split the analysis of the results we describe next into two groups: translation into a TL morphologically richer than the SL (pairs with English as SL), and translation from a morphologically richer SL (pairs with English as TL). It is also worth mentioning that, in all the scenarios evaluated, when a system was trained with interleaved TL tags, the decoder alternately produced TL tags and surface forms at test time as expected.

Translation into a morphologically rich language. When the TL is morphologically richer than the SL, interleaved tags lead to higher BLEU scores, although the impact changes depending on the information encoded in the tag and the language where they are used (SL or TL). SL DUM tags are not very effective: they bring a statistically significant increase in BLEU only to 2 out of the 8 systems evaluated with the recurrent architecture,[7] and to none of the 8 Transformer systems. SL POS and MSD tags generally outperform DUM tags, as they contain information that helps to obtain a better representation of the SL sentence and break the grammatical ambiguity of English (Sennrich and Haddow, 2016). There is a statistically significant difference between SL POS and SL MSD tags in 6 out of the 16 systems evaluated,[8] and in 5 out of these 6 systems MSD tags outperform POS tags. For some language

---

[6]Those systems trained with both SL MSD and TL POS tags could not be compared with systems with both SL DUM and TL DUM tags because the latter were not included in the experimental set-up in order to keep the experiments to a manageable size. Hence, their scores do not contain any bullet.

[7]Four language pairs and two training corpus sizes.

[8]Four language pairs, two training corpus sizes and two architectures.

Language	Tags	English as SL				English as TL			
		recurrent		Transformer		recurrent		Transformer	
		50k	Full	50k	Full	50k	Full	50k	Full
Czech	None (baseline)	7.59	11.98	6.97	12.35	12.34	17.81	10.63	17.58
	SL DUM	7.61	12.30	6.87	12.66	**12.79**	17.98	**11.05**	17.79
	SL POS	7.75	12.33	7.18	12.62	**13.06**	18.05	**11.00**	**17.99**
	SL MSD	**8.15**•†	12.55	**7.35**•	**13.09**•†	**13.39**•	18.08	**11.70**•†	**18.71**•†
	TL DUM	7.64	12.14	6.49	12.75	12.69	17.94	10.78	18.62
	TL POS	**8.11**•	**12.58**•	**7.50**•†	**13.68**•†	**13.38**•	**18.76**•	11.14	18.33
	TL MSD	**8.08**•	12.41	7.14•	12.98	**13.13**•	**18.54**•	**11.23**•	18.45
	SL MSD/TL POS	**9.02**	**13.51**	7.95	13.27	**14.61**	**19.40**	**11.65**	**19.35**
German	None (baseline)	17.17	26.79	15.07	28.32	18.52	27.61	14.97	27.73
	SL DUM	17.19	**27.40**	14.98	27.97	18.85	27.98	15.35	27.90
	SL POS	**18.43**•	**27.78**	**16.34**•	**29.16**•	**19.63**•	27.99	**15.68**	**28.32**•
	SL MSD	**18.41**•	27.52	**16.36**•	29.01•	**18.41**•	27.97	**15.86**•	27.98
	TL DUM	17.77	**27.50**	15.17	27.67	17.77	27.96	15.17	27.67
	TL POS	**17.97**	**28.06**•†	15.48†	**29.07**•†	**17.97**	**28.28**•	**15.48**•†	**29.07**•
	TL MSD	17.53	27.22	14.91	28.34•	17.53	**28.45**•	14.91	28.34
	SL MSD/TL POS	**20.44**	**29.31**	**17.20**	**29.46**	**20.44**	**29.78**	**17.20**	**29.46**
	WMT None (baseline)	-	37.97	-	38.59	-	39.94	-	40.12
	WMT SL MSD/TL POS	-	**39.51**	-	38.48	-	**40.61**	-	40.40
Spanish	None (baseline)	20.75	26.75	19.18	26.78	19.88	25.59	18.03	25.82
	SL DUM	**21.13**	27.03	19.28	26.63	**20.30**	**26.18**	18.46	25.80
	SL POS	**21.88**•	26.92	**19.57**	**27.11**•†	**20.78**•	26.17	**18.72**	25.77
	SL MSD	**21.68**•	**27.29**†	**20.00**•†	26.60	20.49	26.22	18.55	25.82
	TL DUM	20.93	26.79	19.57	25.95	19.86	25.66	18.06	25.72
	TL POS	**21.29**•	27.04	19.60	27.02•	**20.91**•†	**26.17**•	**18.76**•	26.09†
	TL MSD	**21.43**•	26.91	19.72	27.23•	**20.45**•	**26.13**•	**18.82**•	25.60
	SL MSD/TL POS	**22.69**	**27.79**	20.35	26.89	**21.87**	**27.02**	**19.07**	**26.30**
Turkish	None (baseline)	6.21	9.78	4.24	10.30	10.66	15.00	8.30	15.50
	SL DUM	6.14	9.78	4.31	10.60	10.18	15.28	8.24	15.90
	SL POS	**6.98**•	**10.37**•	4.38	**10.69**	10.65•	15.36	8.06	**16.10**
	SL MSD	**7.02**•	**10.80**•†	**4.62**•	**10.69**	**10.73**•	15.39	7.90•	15.93
	TL DUM	5.90	9.96	4.20	10.39	10.37	14.79	8.32	16.21
	TL POS	**6.50**•†	10.08†	4.50•	**10.66**	**10.84**•†	14.56	**9.43**•†	16.12
	TL MSD	5.96	9.63	4.59•	10.48	10.23	**15.90**•†	8.75•	16.42
	SL MSD/TL POS	**7.27**	**10.85**	4.95	10.84	**12.08**	**16.75**	9.38	16.99

Table 2: BLEU scores computed on the test set for the different language pairs, corpus sizes and setups evaluated. See Section 4 (first paragraph) for a description of the different symbols annotating the scores.

pairs and training corpus sizes, enriching the SL representation with information about number, verbal mood, etc. proves to be useful.

We can find stronger differences between the different types of tags in the TL. TL DUM tags are useful for the recurrent English–German systems, in line with the findings by Wagner (2017), but their contribution to other language pairs and the Transformer architecture is less clear. The most relevant trend is that using only POS tags in the TL consistently outperforms the use of MSD tags: statistically significant differences are found in all TLs but Spanish, a Romance language which has the simplest morphology. This result is further investigated in sections 5 and 6. Finally, combining SL MSD and TL POS tags leads to the highest scores.

Translation from a morphologically rich language. The effects of using SL tags when the SL is morphologically richer than the TL are similar to those observed in the opposite scenario: POS and MSD tags often outperform DUM tags. When statistically significant differences between POS and MSD tags are found, they favour MSD tags. Concerning TL tags, the systematic degradation observed for MSD tags is less frequent than in the opposite direction, and it is mainly concentrated in the smallest corpus size. An explanation could be that morphological information in English is less complex and easier to predict from the SL sentence. Finally, combining SL MSD tags and TL POS tags also leads to the highest scores.

Large-scale training data. The results for the English–German WMT large-scale training data, also depicted in Table 2, show a different picture. We can still observe that the use of interleaved tags brings a statistically significant improvement, but this only happens in the recurrent architecture. Transformer
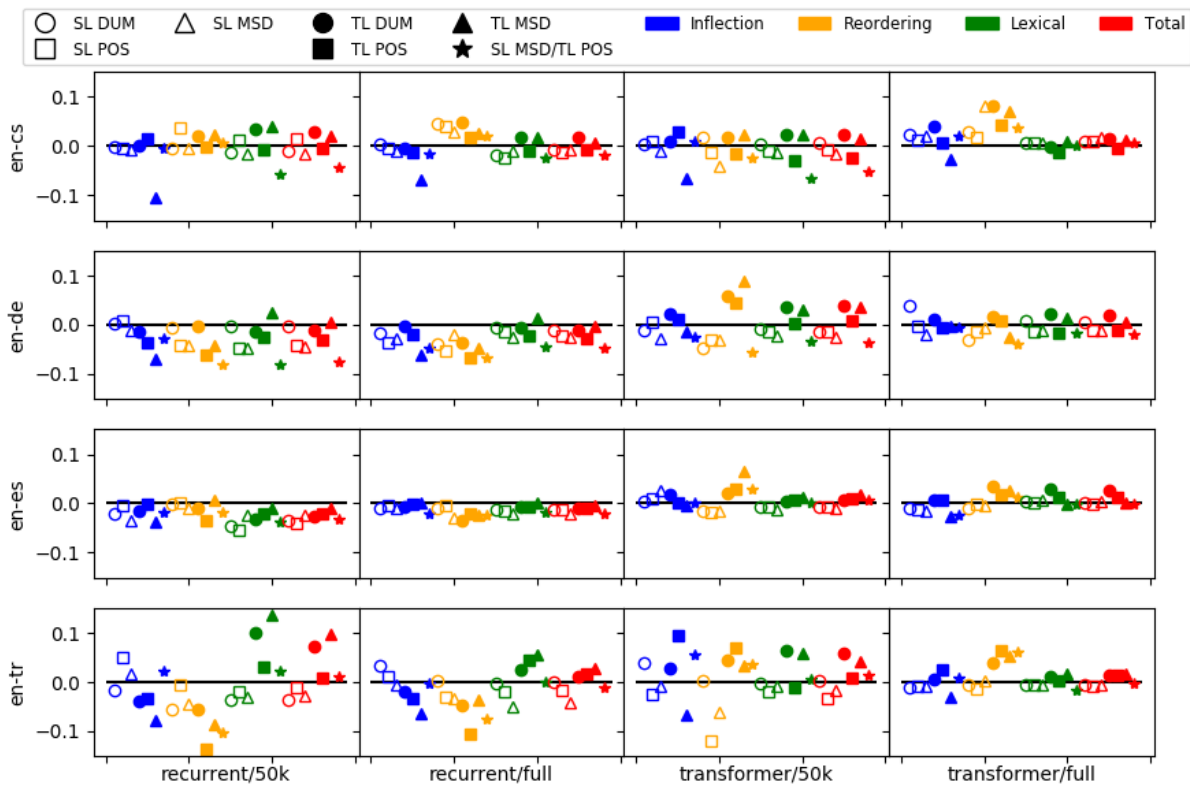
Figure 1: For language pairs with English as SL, relative changes in the number of errors for each error category, training corpus size and type of interleaved tag.

systems do not benefit from the interleaved linguistic annotations when the training corpus size is large.[9] A potential explanation is provided in the next section.

Main findings. In line with previous works (Nadejde et al., 2017; Wagner, 2017), the results analysed so far suggest that interleaved linguistic annotations are helpful both in the SL and the TL and they should be included in both languages in order to maximize performance. While morphological features can be useful in the SL, they should be avoided in the TL if it is morphologically rich. Even when large corpora are available, linguistic annotations can help to boost translation quality.

5 Error analysis

To better understand the results obtained, we computed the relative difference in the number of Hjerson errors between the systems with interleaved tags and the baseline;[10] a positive value means that the system made more errors than the baseline. As we did before, we split the results into two groups of language pairs: those with English as SL, depicted in Figure 1, and those with English as TL, depicted in Figure 2. In the remainder of this section, we analyse the results obtained and illustrate them with examples.

SL tags. SL POS and SL MSD tags systematically reduce lexical errors (green, empty squares and triangles are below the horizontal line). Reordering errors are also reduced with the exception of English–Czech, in which the TL has a relatively flexible word order.[11] Concerning inflection errors, there is not

---

[9]In order to confirm that the differences observed were not caused by the random initialisation of the network weights, we trained the systems on WMT data two more times. The BLEU scores obtained for the additional training runs confirm that linguistic annotations only help in recurrent systems. We did not perform additional training runs for the under-resourced scenario because of time constraints and because the conclusions drawn for that scenario are based on the results obtained for different language pairs (and hence training runs), which we think makes them solid enough.

[10]Computed as (#errors_interleaved − #errors_baseline)/#errors_baseline.

[11]Improvements in reordering quality are not captured by the automatic error classification if translation hypotheses with an acceptable word order that differs from that in the reference are produced by the system.
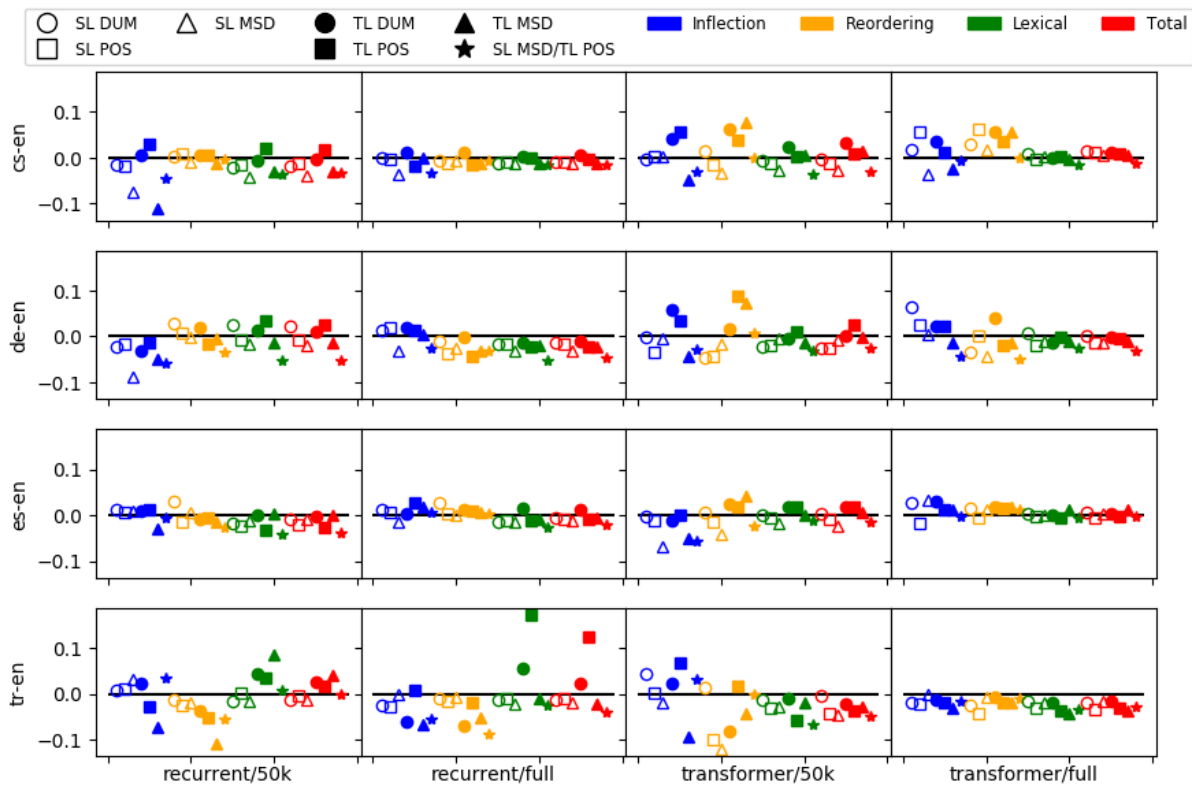
Figure 2: For language pairs with English as TL, relative changes in the number of errors for each error category, training corpus size and type of interleaved tag.

clear trend. As aforementioned, a possible explanation could be that SL tags help to obtain more accurate representations of the SL sentences; since inflection errors are related to modeling TL grammar rather than to representing the SL sentence, they are not reduced by interleaving SL tags. All these results are compatible with the evaluation metrics, which showed that SL tags generally improve translation quality. In the first example in Table 3, SL tags help to obtain a better representation of the SL sentence: the system is able to interpret that *matters* is acting as a noun and produces *hace que las cosas sean simples* (en: *it makes things simple*) instead of *hace que los alemanes sean simples* (en: *it makes Germans simple*).

TL tags. Different error distributions can be observed depending on the information encoded in the TL tags. TL MSD tags systematically reduce inflection errors in both architectures (the blue, filled triangle is usually among the lowest points in the figure). The largest inflection error reductions occur with highly inflected TLs such as Czech and Turkish. TL POS tags, on the contrary, do not systematically reduce inflection errors. Hence, the system using TL MSD tags is using the morphological features they encode

Pair	Tags	Sentence
en-es	orig/ref.	This does not make **matters simple** for the Germans. → Esto no resulta fácil para los alemanes.
	None	Esto no hace **que los alemanes sean simples**.
	SL MSD	Esto no hace **que las cosas sean simples** para los alemanes.
en-de	orig/ref,	Miller was not engaging in literary criticism → Miller engagierte sich nicht in **literarischer** Kritik
	TL POS	Miller setzte sich nicht an **literarische** Kritik
	TL MSD	Miller befasste sich nicht mit **literarischer** Kritik
en-de	orig/ref	"The children should help prepare the food." → "die Kinder sollen **helfen**, das Essen zuzubereiten".
	TL POS	"die Kinder sollten dabei **helfen**, die Nahrung vorzubereiten".
	TL MSD	"die Kinder sollten dazu **beitragen**, die Lebensmittel vorzubereiten"

Table 3: Examples of how interleaving SL or TL tags change the translations. Translations obtained with the recurrent systems trained on the full News Commentary corpus.
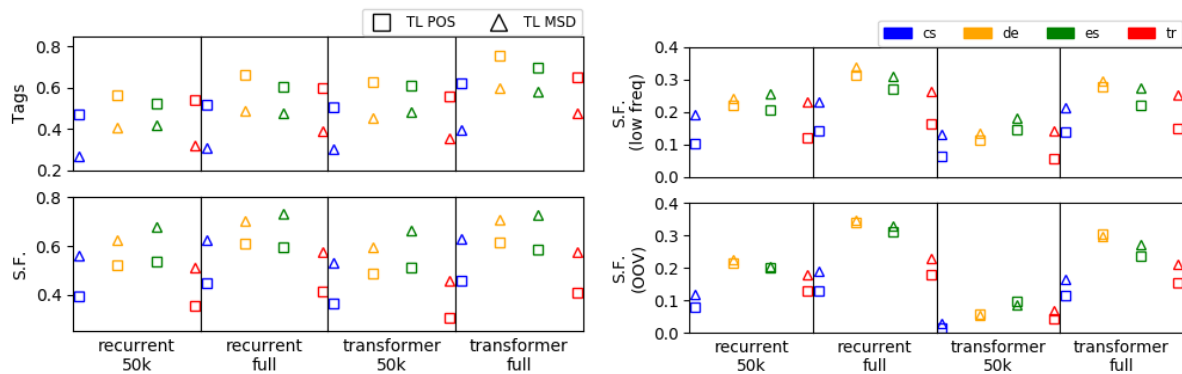
Figure 3: For language pairs with English as SL, tag prediction accuracy (labelled as *Tags*) and surface form prediction accuracy (labelled as *S.F.*) forcing, respectively, surface forms and tags from the reference.

(tense, number, etc.) for producing the correct inflected form according to the reference. In the second example in Table 3, the system using TL MSD tags generates the right inflected form of the German word *literarische* because it has predicted its dative case first.

However, the prediction of MSD tags with complex morphology (see Figure 1) also leads to an increase in lexical errors in comparison with the prediction of POS tags. It can be observed that TL MSD tags bring an increase in lexical errors over the baseline (note the green, filled triangles at the top of the figures), while the impact in lexical errors of the POS tags is less clear. Similarly to inflection errors, the difference between the increases of lexical errors brought by MSD and POS tags is larger for Turkish and Czech, which are the two languages with the most sparse MSD tags. Turkish is a Turkic agglutinative language and Czech is a Slavic fusional language with seven cases and four genders. This is compatible with the automatic evaluation metrics: although using TL MSD tags leads to a more grammatical output, the increase in lexical errors makes the system produce translations that are overall less similar to the reference. Note that lexical errors are the most frequent ones.[12] The third example in Table 3 shows that the system with TL MSD tags translates the verb *help* as *beitragen* rather than *helfen*, which is a more precise translation in that context.

When SL MSD tags and TL POS tags are both interleaved, there is a general reduction in the three error categories as compared with the systems using only tags in one of the languages. This confirms that the advantages of SL and TL tags are complementary.

Differences between architectures. Finally, there is a noticeable difference in how the type of errors made by the systems change when interleaving TL tags in recurrent and Transformer architectures. Reordering errors consistently increase in Transformer systems, while they tend to decrease in recurrent systems. Moreover, TL DUM tags consistently increase the total number of translation errors when they are added to a Transformer system and the TL is highly inflected (observe the red, filled circle usually above the horizontal line in Figure 1), while their impact is not clear in recurrent systems. These two findings suggest that adding extra tokens to the TL stream is not the best way of introducing linguistic annotations in self-attention-based NMT systems. It could also explain the results for the large-scale WMT data, where only recurrent systems were able to take advantage of the linguistic annotations. This hypothesis is also compatible with the results reported on WMT data by Yang et al. (2019), who successfully leveraged TL linguistic annotations in Transformer systems using an *ad-hoc* architecture.

6 Analysing the effect of target language morphology

We compare the output of the systems interleaving TL POS and TL MSD tags in order to ascertain whether the increase in lexical errors is caused by the difficulty of predicting the more complex and sparse MSD

---

[12]For the baseline recurrent system, the average and standard deviation over the four language pairs with English as SL of the absolute number of errors of each type in the test set, expressed in thousands, are as as follows. Inflection: 4.1 ± 0.7, Reordering: 5.3 ± 0.6, Lexical: 28.7 ± 2.1.

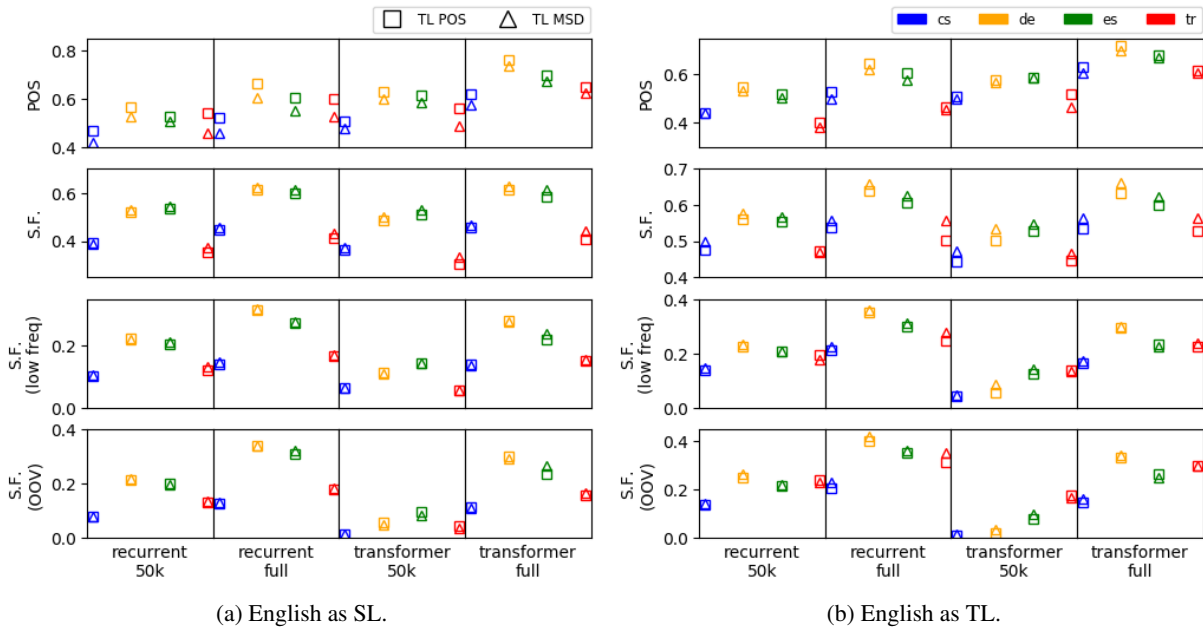(a) English as SL.

(b) English as TL.

Figure 4: POS prediction accuracy (labelled as *POS*) and surface form prediction accuracy (labelled as *S.F.*) forcing, respectively, surface forms and POS tags from the reference.

tags, or by the conditioning of the prediction of surface forms on TL MSD tags. In the latter case, there is the risk that the system learns to strongly condition on tags and avoids generating new words (Tamchyna et al., 2017). This problem could be exacerbated by the sparsity of TL MSD tags because some of them may co-occur only with a few surface forms in the training corpus.

We tried to answer this question by independently evaluating the prediction of tags and surface forms, and comparing the systems interleaving POS and MSD tags. The prediction of surface forms was evaluated by re-decoding the test set and forcing the system to choose the tags from the reference during beam search, whereas the prediction of tags was evaluated by forcing the surface forms from the reference. If the system really learned to strongly condition on tags, when a tag was observed together with only a few surface forms in the training corpus, low-frequency words would not be generated when translating the test set. To test this hypothesis we studied the surface form prediction accuracy for two subsets: infrequent words (frequency ≤ 10 in the training set) and out-of-vocabulary (OOV) words. Figure 3 shows the results for those language pairs with English as SL. It can be observed that there is a trade-off between tag and surface form prediction accuracy: MSD tags are more difficult to predict, but conditioning on them leads to better surface form prediction. On low-frequency and OOV words, MSD tags still outperform POS tags in terms of surface form prediction accuracy, although the difference between them is smaller.

For a fair comparison of both types of tags, we computed the part-of-speech prediction accuracy when predicting MSD tags. The results are depicted in the rows labelled as *POS* in Figure 4. For highly inflected TLs, those systems that predict MSD tags have consistently lower POS prediction accuracy than those that predict only POS tags. The difference is larger for recurrent systems. These results suggest that the difficulty of predicting together the part of speech and its morphological features is indeed one of the reasons behind the lexical degradation of systems using MSD tags. Sparseness of MSD tags seems to play an important role in this degradation: highly inflected TLs present the largest degradation.

To evaluate the impact of predicting morphological features regardless of the low part-of-speech accuracy of MSD tags, we re-computed surface form prediction accuracy by letting the beam search algorithm choose among those MSD tags with the part of speech in the reference. The results, depicted in the rows labelled as *S.F.* in Figure 4, show that, if the systems with interleaved MSD tags correctly predicted the part of speech, the surface form predictions would not be worse than those of systems with interleaved POS tags, neither in general nor for low-frequency and out-of-vocabulary words. Hence,

errors in surface form prediction arising from strongly conditioning on sparse MSD tags do not seem to be the main cause behind the degradation of translation quality brought by TL MSD tags. Actually, when tags with the part of speech of the reference are chosen, conditioning on MSD tags outperforms conditioning on POS tags in terms of overall surface form prediction accuracy for language pairs with English as TL. For the other language pairs, the gain introduced by MSD tags is less clear. One possible reason could be that BPE segmentation does not allow the system to learn a general mapping between tags and word endings from the training data. Another explanation could be related to the fact that predicting the morphological gender for German, Czech and Spanish forces the tag prediction task to be aware of TL lexical information, preventing an optimum division of labour between tag and surface form predictions.

In conclusion, the prediction of TL morphological information needs to be factorised differently in order not to harm part-of-speech prediction. For instance, in the morphological analysis field, Chaudhary et al. (2019) and Straka et al. (2019) predict part of speech and each morphological attribute independently.

7 Concluding remarks

In this paper, we have studied the effects of using linguistic annotations of SL and TL words in under-resourced NMT by interleaving linguistic tags for different language pairs, architectures, training data sizes and types of linguistic information (part of speech and morpho-syntactic descriptions).

We have shown that both SL and TL linguistic annotations are useful, in line with previous works in the literature (Wagner, 2017). SL linguistic annotations lead to more accurate SL sentence representations, and for some language pairs, the use of morpho-syntactic descriptions (consisting of part of speech and morphological features) improves the representation obtained when only part-of-speech tags are used. Surprisingly, for highly inflected TLs, TL linguistic annotations are more useful if they simply consist of part-of-speech information. Using morpho-syntactic descriptions leads to an overall translation quality degradation in terms of automatic evaluation metrics, even though it improves the grammaticality of the output. We have also shown that predicting TL morpho-syntactic descriptions frequently results in wrong part-of-speech predictions. Hence, to optimize the use of TL morphological information in NMT, it is advisable to avoid the prediction of part-of-speech and morphological features together as monolithic tags.

The gain introduced by linguistic information encoded as interleaved tags scales to large data availability scenarios only for the recurrent architecture. This result, together with the conclusions of the automatic error analysis, suggest that adding extra tokens to the TL stream is not the optimum way of introducing additional linguistic information in self-attention-based NMT systems.

In summary, the use of morpho-syntactic descriptions in the SL and part-of-speech information in the TL, which can be easily obtained even in under-resourced scenarios, systematically improves translation quality when they are simply interleaved in training data as linguistic tags, even without using a morphological generator (Tamchyna et al., 2017), which could be error-prone for under-resourced languages, and without any kind of information about syntactic structures.

Acknowledgements

References

Aditi Chaudhary, Elizabeth Salesky, Gayatri Bhat, David R. Mortensen, Jaime Carbonell, and Yulia Tsvetkov. 2019. CMU-01 at the SIGMORPHON 2019 shared task on crosslinguality and context in morphology. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 57–70, Florence, Italy, August. Association for Computational Linguistics.

Kehai Chen, Rui Wang, Masao Utiyama, Lemao Liu, Akihiro Tamura, Eiichiro Sumita, and Tiejun Zhao. 2017. Neural machine translation with source dependency representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2846–2852, Copenhagen, Denmark, September. Association for Computational Linguistics.

Xiaocheng Feng, Zhangyin Feng, Wanlong Zhao, Nan Zou, Bing Qin, and Ting Liu. 2019. Improved neural machine translation with pos-tagging through joint decoding. In Shuai Han, Liang Ye, and Weixiao Meng, editors, *Artificial Intelligence for Communications and Networks*, pages 159–166, Cham. Springer International Publishing.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the 13th International Workshop on Spoken Language Translation*, December.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2017. Extending hybrid word-character neural machine translation with multi-task learning of morphological analysis. In *Proceedings of the Second Conference on Machine Translation*, pages 296–302, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jetic Gū, Hassan S. Shavarani, and Anoop Sarkar. 2018. Top-down tree structured decoding with syntactic connections for neural machine translation and parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 401–413, Brussels, Belgium, October-November. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Conference Track Proceedings of the 3rd International Conference on Learning Representations*, San Diego, USA.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, volume 4, pages 388–395, Barcelona, Spain.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 12–18, Florence, Italy, August. Association for Computational Linguistics.

Qiang Li, Derek F. Wong, Lidia S. Chao, Muhua Zhu, Tong Xiao, Jingbo Zhu, and Min Zhang. 2018. Linguistic knowledge-aware neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(12):2341–2354, December.

Maria Nadejde, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra 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. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Comput. Linguist.*, 37(4):657–688, December.

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

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher 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. Association for Computational Linguistics.

Rico Sennrich and Barry 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. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany, August. Association for Computational Linguistics.

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, August. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark, September. Association for Computational Linguistics.

Mark Steedman. 2000. *The syntactic process*, volume 24. MIT press Cambridge, MA.

Milan Straka, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy, August. Association for Computational Linguistics.

Aleš Tamchyna, Marion Weller-Di Marco, and Alexander Fraser. 2017. Modeling target-side inflection in neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 32–42, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Antonio Toral and Víctor 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. Association for Computational Linguistics.

Francis M Tyers and Murat Serdar Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Martin Wagner. 2017. Target factors for neural machine translation.

Xinyi Wang, Hieu Pham, Pengcheng Yin, and Graham Neubig. 2018. A tree-based decoder for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4772–4777, Brussels, Belgium, October-November. Association for Computational Linguistics.

Shuangzhi Wu, Dongdong Zhang, Zhirui Zhang, Nan Yang, Mu Li, and Ming Zhou. 2018. Dependency-to-dependency neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 26(11):2132–2141, November.

Xuewen Yang, Yingru Liu, Dongliang Xie, Xin Wang, and Niranjan Balasubramanian. 2019. Latent part-of-speech sequences for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 780–790, Hong Kong, China, November. Association for Computational Linguistics.

Hao Zhou, Zhaopeng Tu, Shujian Huang, Xiaohua Liu, Hang Li, and Jiajun Chen. 2017. Chunk-based bi-scale decoder for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 580–586, Vancouver, Canada, July. Association for Computational Linguistics.

Language	English as SL		English as TL	
	50k	Full	50k	Full
Czech	5 000;False;512; 512	10 000;True;1024; 512	5 000;False;1024; 512	10 000;False;1024; 512
German	5 000;True;1024; 512	10 000;True;1024; 512	5 000;False;1024; 512	10 000;False;1024; 512
Spanish	5 000;False;1024; 512	5 000;True;1024; 512	5 000;True;1024; 512	5 000;False;1024; 512
Turkish	5 000;False;1024; 512	10 000;False;1024; 512	5 000;True;1024; 512	10 000;False;1024; 512

Table 4: Optimum hyper-parameters for recurrent models. Each cell contains, in this order: number of BPE operations, whether TL input and output embeddings are shared, size of the hidden layer and size of the embedding layer.

Language	English as SL		English as TL	
	50k	Full	50k	Full
Czech	20 000;True;256; 4; 4	10 000;True;256; 4; 4	20 000;True;256; 4; 4	20 000;True;256; 4; 4
German	5 000;True;256; 4; 4	10 000;True;256; 4; 4	20 000;True;256; 4; 4	10 000;True;256; 4; 4
Spanish	5 000;True;256; 4; 4	10 000;True;512; 6; 8	5 000;True;256; 4; 4	10 000;True;512; 6; 8
Turkish	20 000;True;128; 2; 2	10 000;True;256; 4; 4	5 000;True;128; 2; 2	20 000;True;256; 4; 4

Table 5: Optimum hyper-parameters for Transformer models. Each cell contains, in this order: number of BPE operations, whether TL input and output embeddings are shared, size of the model, number of layers and number of attention heads.

A Training details

The optimum training hyper-parameters were obtained by following the grid search process depicted next. At each step, we chose the hyper-parameters that maximised BLEU on the development set. Table 4 shows the optimum hyper-parameters for each language pair and training corpus size for the recurrent architecture while Table 5 shows the same information for the Transformer architecture.

- First, we explored the optimum number of BPE operations among the following values: 5,000, 10,000, 20,000, and 40,000. The rest of hyper-parameters were set to the values recommended by Sennrich et al. (2017) for the recurrent architecture and by Vaswani et al. (2017) for the Transformer architecture ("base" configuration), respectively.

- With the optimum number of BPE operations, we then tested if better results could be obtained with tied embeddings (Press and Wolf, 2017) in the decoder.

- Finally, we explored the following combinations of hyper-parameter values for each architecture with the best number of BPE operations and tied embedding configuration obtained in the previous steps.

  recurrent: Hidden and embedding sizes: $(1024, 512)$, $(512, 512)$, $(512, 256)$ and $(256, 256)$.

  Transformer: Model size, number of layers in the encoder and the decoder and number of attention heads: $(512, 6, 8)$, $(256, 4, 4)$, $(128, 2, 2)$.

For all systems trained, we applied label smoothing with a value of 0.1 and dropout of 0.1. Unlike Sennrich and Zhang (2019), we did not use a lexical model neither word dropout. The optimisation algorithm was Adam (Kingma and Ba, 2015) with the inverse square root learning rate decay (Vaswani et al., 2017, Sec. 5.3) and 8,000 warm-up iterations. Learning rates were initialised to 0.0004 for recurrent and to 0.0003 for Transformer. Training stopped after 10 validations without any perplexity improvement on the development corpus; validations were performed every 1,000 mini-batches. The model finally used is the one for which the best BLEU score was obtained on the development corpus. We used the same amount of sentences per mini-batch for all the models trained for a given TL and corpus size; the amount of sentences in each mini-batch ensures that, when SL and TL MSD tags are interleaved, the amount of tokens is below 4,500.