Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D2.2 – GoURMET Final report on modelling morphological structure

Nature	Report	**Work Package**	WP2
Due Date	30/06/2022	**Submission Date**	30/06/2022
Main authors	Barry Haddow (UEDIN), Víctor M. Sánchez-Cartagena (ALAC), Wilker Aziz (UvA)		
Co-authors			
Reviewers	Jindřich Helcl		
Keywords	morphology, machine translation		
Version Control			
v1.0	**Status**	Final	28 June 2022

Contents

List of Figures

Abstract

In this deliverable we describe the research of WP2 (Modelling Morphological Structure) in the second half of the GoURMET project. The idea of this work package was to understand how the morphological structure of words could be best exploited to improve low-resource machine translation. The research consisted of four pieces of work: (i) we examined how word-level linguistic annotation can be employed on the target-side, in order to improve MT.; (ii) We developed a pipeline to reproduce state-of-the-art morphological segmentation and lemmatisation, which we release as open-source; (iii) We studied the effect of different segmentation algorithms on the translation of agglutinative languages; and (iv) we examined how measures of morphological typology could be used to predict difficulty in MT.

1 Introduction

This deliverable describes the work done in WP2 (Modelling Morphological Structure) in the second half of the GoURMET project.

From the grant agreement, the aim of this work package is as follows:

The purpose of this work package is to develop methods for representing words within neural machine translation systems. The techniques developed in this work package will specifically target the low resource setting. WP2 has three main goals:

- *Develop methods which rely on prior knowledge to induce models of linguistically-plausible morphological structure (Task 2.1)*
- *Develop methods to induce alignments and morphological structure jointly (Task 2.2)*
- *Develop translation models which induce implicit word structure as latent features (Task 2.3)*

We describe the work under each of the tasks in the following sections. In Task 2.1 we carried out a study of word-level linguistic annotation in the target language (Section 2.1) as well as developing a pipeline to reproduce state-of-the-art morphological analysis and lemmatisation (Section 2.2). For Task 2.2 we investigated how different subword strategies coped with translation between agglutinative languages (Section 3.1) whilst for Task 2.3 we looked at how measures of morphological typologies could be used to predict difficulty in MT (Section 4.1).

2 Linguistically informed models of morphology

2.1 Word-level linguistic annotations

In under-resourced scenarios, the use of additional information in the form of linguistic annotations at the word level (such as part-of-speech, morphological or syntactic tags) can improve neural machine translation (NMT) performance (Sennrich and Haddow, 2016; Nadejde et al., 2017). Linguistic annotations may be integrated in the source language (SL), where they help the model to produce more accurate SL representations (Sennrich and Haddow, 2016); or in the target language (TL), where their addition involves producing probability distributions for both TL words and TL linguistic annotations (García-Martínez et al., 2016).

We previously conducted an exhaustive study about the effect of word-level linguistic annotations in under-resourced NMT using part-of-speech (POS) tags and morpho-syntactic description (MSD) tags in both the SL and the TL (Sánchez-Cartagena et al., 2020, see also Sect 2.2 of deliverable D2.1). In that study, we simply *interleaved* (Nadejde et al., 2017) linguistic annotation tags in the input and output streams. Our results showed that, overall, in the SL MSD tags are a better choice than POS tags, whereas in the TL POS tags outperform MSD tags. Apparently, learning the additional correlations needed to produce MSD tags in the TL is more difficult than just learning the part-of-speech.

In the piece of work being presented in this deliverable, we delve further into our previous study to shed light on the contradictory evidence found in the literature (Wagner, 2017; Feng et al., 2019) as regards the best way of combining the generation of linguistic annotations and surface forms in the TL. On the one hand, as in the aforementioned interleaving approach, the generation of a TL surface form can be explicitly conditioned on its corresponding tag. In this way, TL tags and surface forms are generated in alternate time steps and the probability of a TL sentence is factorized as follows, being y the sequence of surface forms and t the corresponding sequence of linguistic tags:

$$\prod_{i=1}^{|y|} p(y_i|t_{1..i}, y_{1..i-1}, x; \lambda_w) \cdot p(t_i|t_{1..i-1}, y_{1..i-1}, x; \lambda_t) \tag{1}$$

On the other hand, the generation of linguistic tags could be used to enrich the representations from which the surface forms are generated, thus avoiding the generation of linguistic tags at decoding time (García-Martínez et al., 2016). In this approach, the probabilities of TL surface forms and linguistic tags are conditionally independent, and the effect of the TL tags is achieved via parameter sharing. This approach is usually referred to as multi-task learning (MTL) in the literature, since the probability of a TL sentence is factorized as follows, being y the sequence of surface forms and t the corresponding sequence of linguistic tags:

$$\prod_{i=1}^{|y|} p(y_i|y_{1..i-1}, x; \lambda_w) \cdot \prod_{i=1}^{|y|} p(t_i|y_{1..i-1}, x; \lambda_t) / \lambda_w \cap \lambda_t \neq \emptyset \tag{2}$$

In both approaches, the training objective is usually the categorical cross-entropy loss. Hence, the training objective for interleaving becomes is the following, where y the gold-standrard sequence of surface forms and t its corresponding sequence of linguistic tags:

$$\sum_{i=1}^{|y|} \log p(y_i|t_{1..i}, y_{1..i-1}, x; \lambda_w) + \log p(t_i|t_{1..i-1}, y_{1..i-1}, x; \lambda_t) \tag{3}$$

Similarly, the training objective for MTL is:

$$\sum_{i=1}^{|y|} \log p(y_i|y_{1..i-1}, x; \lambda_w) + \sum_{i=1}^{|y|} \log p(t_i|y_{1..i-1}, x; \lambda_t) \tag{4}$$

Moreover, for each of the two approaches, there are multiple potential strategies for sharing the network parameters between the two tasks, and most of them remain unexplored. For the alternate generation approach, the way of sharing parameters commonly found in the literature (Sánchez-Cartagena et al., 2020) is using exactly the same set of parameters (i.e. exactly the same NMT system) for both outputs. However, since it is clear that, for a certain time step, the system must emit either a linguistic tag or a surface form, mixing them in the same probability distribution (generated by a softmax operation) does not seem to be the most appropriate way of modelling the sequence generation. One could split the last linear layer of the model, and use a different one for each type of output. This separation of parameters could be further extended to previous layers, and even to the full set of decoder parameters.

For the so-called MTL approach, there are also multiple ways of sharing the parameters. In order to share all parameters, one could train the system as if it was a one-to-many multilingual system (Johnson et al., 2017), where the linguistic tags are regarded as one of the target languages. Another option, as proposed by García-Martínez et al. (2016), is to emit both the surface form and the linguistic tag from a common representation layer at each time step. This common representation is usually the output of the network before the last linear layer; it can be however further reduced to build a system with an independent decoder for each task.

With the aim of assessing the performance of the approaches and parameter sharing strategies that have just been described, we carried out a set of experiments with well-known low-resource datasets for Korean–English and German–Upper Sorbian. We also trained English–German systems with decreasing amounts of training data. Preliminary results for these experiments are described next.

In line with our previous findings (Sánchez-Cartagena et al., 2020, Sec. 6), the results suggest that the additional correlations that need to be learned in systems with alternate generation of tags and surface forms increase data sparseness and harm translation quality when training data is scarce (around 1 million words per language). In these scenarios, MTL seems to be a more robust alternative. MTL is also more robust against errors in the linguistic annotation, likely to happen in under-resourced languages. On the contrary, when the training corpus size grows, interleaving generally outperforms MTL.

Concerning the different parameter sharing strategies evaluated, sharing all the parameters seems to be the best strategy when tags and surface forms are generated in an interleaved way. In the MTL approaches, using different linear layers, as proposed by García-Martínez et al. (2016), is the most effective strategy. However, it requires an appropriate down-weighting of the linguistic tag loss with regard to the surface form loss. Using independent decoders was the least performing sharing strategy for both approaches.

2.2 Lemmatisation and Morphological Tagging

This section reports on an engineering effort to reproduce state-of-the-art systems for morphological lemmatisation and tagging based on data from the CONLL/SIGMORPHON 2019 shared task 2 (McCarthy et al., 2019).

State-of-the-art tagging with UDPipe2. Designed and implemented by Straka et al. (2019) at Charles University, the second version of the UDPipe pipeline (Straka, 2018) has shown excellent performance in several competitions. It consists of an entirely modular system (see Figure 1), processing various forms of pre-trained (e.g., fasttext (Bojanowski et al., 2016)) and trainable word embeddings (e.g., regular embedding layers and character-level bidirectional GRU encoders) via a residual RNN, before classifying a tokens' morphological tag set and lemma edit script. Ultimately, this UDPipe2 was 1 of 3 winners at CONLL 2018's shared task, and 1 of 2 at CONLL/SIGMORPHON 2019 shared task 2 (McCarthy et al., 2019), and the winner at the EvaLatin 2020 shared task (Sprugnoli et al., 2020).

Reimplementation. We are motivated by the need for accurate morphological analysis for multiple languages and the fact that we did not have access to a state-of-the-art system, such as UDPipe or UDPipe 2 (Straka, 2018; Straka et al., 2019). Besides, since the CONLL/SIGMORPHON 2019 shared task 2 (McCarthy et al., 2019) took place, the universal dependencies (UD) treebanks have seen 6 new releases,[1] with improvements and extensions made to many of the used corpora. To use

[1] https://universaldependencies.org

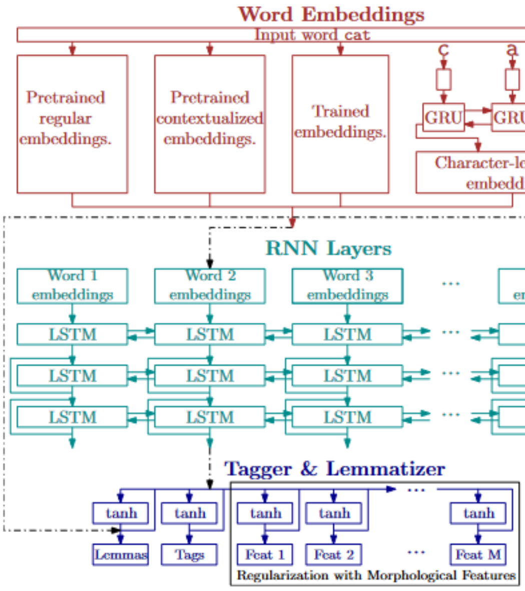

Figure 1: UDPipe2's pipeline. Image from (Straka et al., 2019).

these datasets instead, we have converted UD2.9 to carry UniMorph tags (McCarthy et al., 2018) using the ud-compatibility repo.[2] We strived for a well designed pipeline that employs modern components and delivers reasonably fast tagging performance (measured in tokens per second) across a large set of languages. See Table 1 for test results of our reimplementation. The open-source code is available on https://github.com/IvoOVerhoeven/morph_tag_lemmatize and it will continue to be improved.

Language	Lemma		Morphology			Tokens/sec
	acc. ↑	Lev. ↓	set acc. ↑	tag micro F1 ↑	tag macro F1 ↑	
Arabic	0.93	0.21	0.90	0.96	0.85	2313.08
Czech	0.98	0.03	0.92	0.98	0.90	2930.30
Dutch	0.94	0.12	0.95	0.97	0.93	3222.71
English	0.97	0.05	0.92	0.96	0.90	2976.70
Finnish	0.82	0.44	0.81	0.92	0.62	2632.62
French	0.98	0.04	0.92	0.97	0.87	3715.83
Russian	0.97	0.06	0.92	0.97	0.88	2759.40
Turkish	0.91	0.19	0.77	0.89	0.58	1828.43

Table 1: Performance of our taggers across languages.

Potential. In an ongoing study, these taggers are used to annotate the output of NMT models througout training. Based on whether the output is inflected as expected (by comparison of its morphological features—as predicted by our taggers—and those of the reference), we guide a curriculum of training examples aimed at improving a standard NMT system along the morphological

[2] https://github.com/unimorph/ud-compatibility

inflection axis. Given the current relevance of multilingual pre-trained architectures, the ability to inform the NMT system through its curriculum, rather than through changes to its architecture, is particularly appealing, as it is compatible with the pre-train then fine-tune approach (Tang et al., 2021). This project started in GoURMET's last semester in response to the success of massive multilingual pre-trained models, and we will continue to pursue this direction.

3 Jointly learning alignments and morphology

3.1 Investigating Subword Segmentation Strategies and Agglutinative Languages

3.1.1 Introduction

The current state of machine translation is heavily English-centric, which means improvements yielded by certain methods towards the type of morphology English exhibits (i.e. low morphological complexity and fusional), might not transfer when either the source and/or target exhibits a different type of morphology. The more morphologically rich a language is, the greater the data sparsity problem becomes for that language, and by extension, the subword choice or handling may become more imperative. Agglutinative languages are those where words can be typically composed of several morphemes concatenated together to form one word. There has been little research into machine translation between agglutinative languages (as opposed to MT between agglutinative languages and English). For this reason, we decided to test recently proposed segmentation methods which claimed to improve on BPE (Sennrich et al., 2016). We were particularly interested in a new method for finding the optimal merge count for BPE (Xu et al., 2021) – VOLT (vocabulary learning via optimal transport).

3.1.2 Experiments

We decided to compare the following methods which deal with subword choice/regularisation, and see whether the gains are still present in a low-resource agglutinative–agglutinative setting. These methods are diverse in their approach:

- Conventional BPE baseline (Sennrich et al., 2016) with merge count of 2k. The motivation for this merge choice was as recommended by Ding et al. (2019) for a low-resource scenario.

- BPE Dropout (Provilkov et al., 2020). A regularisation method simple to implement, which they claim works best in a LR scenario. Essentially, when applying the BPE merge rules, there is a dropout probability ($p = 0.1$) of the rule *not* being applied. This dropout is applied *on the fly*, therefore, a sentence may be segmented multiple different ways throughout training. We set the same merge count as conventional BPE at 2k.

- Vocabulary Learning via Optimal Transport (VOLT) (Xu et al., 2021). An entropy and information theory based method that aims to select the best BPE merge/vocab size, using a metric the authors define as 'Marginal Utility of Vocabularization'. VOLT is described in more detail below.

All models were transformers[3] (Vaswani et al., 2017) trained using the `fairseq` toolkit (Ott et al., 2019).

Language Pairs of Interest We decided to experiment with languages from a diverse range of families. The following agglutinative language pairs are of natural interest whether for geographic, political or cultural reasons:

- **Uralic:** Finnish–Hungarian and Estonian–Hungarian.
- **Turkic:** Turkish–Kazakh.
- **Dravidian:** Tamil–Telugu and Tamil–Malayalam
- **Bantu:** Swahili–Xhosa.

We also decided to look at each of the languages paired with English, as this would be closer to previous work where English has nearly always been either source or target. Our data setup was to use 200k sentences from the multilingual JW300 corpus (Agic and Vulic, 2019). For development and test set, we used FLORES (Goyal et al., 2021).

VOLT The main motivation of VOLT (Xu et al., 2021) was to find the best vocabulary/subword segmentation of a training corpus for machine translation, without having to do an expensive hyperparameter search (on a parameter such as number of BPE-merges), which is costly in terms of computational resources to train the neural models. They define a metric called "Marginal Utility of Vocabularization" (MUV) as the negative derivative of entropy (of the training corpus, both source and target) with respect to vocabulary size, and aim to maximise this measure using an Optimal Transport technique.

We will use the same notation as that of Xu et al. (2021) throughout our discussion. VOLT works by iterating through sets of different sized vocabularies $\mathbb{V}_{S[t-1]}$, $\mathbb{V}_{S[t]}$, where $\boldsymbol{S} = \{k, 2 \cdot k, ..., (t-1) \cdot k, ...\}$ is an incremental integer sequence with some step size k. Briefly, it proposes to calculate the marginal entropy between each consecutive set, and chooses the vocabulary size that yields the largest entropy difference, using optimal transport[4]. After determining which subwords should be in the vocabulary, VOLT uses the same greedy approach of BPE to merge individual units together.

The authors of VOLT tested their method on a large variety of language pairs, including some low-resource settings. We had hoped to replicate their promising results, although this turned out differently than expected.

We used the implmentation of VOLT provided by the authors. In order to obtain our VOLT vocabulary, we performed the following; after tokenising the datasets, we learnt the joint-BPE up to 10,000 merges. We set k to be 400 (the vocabulary step size) i.e. our $\boldsymbol{S} = \{400, 2 \cdot 400, ..., (t-1) \cdot k, ..., 10,000\}$. After the desired vocabulary size was chosen, we used those merge rules to segment the data.

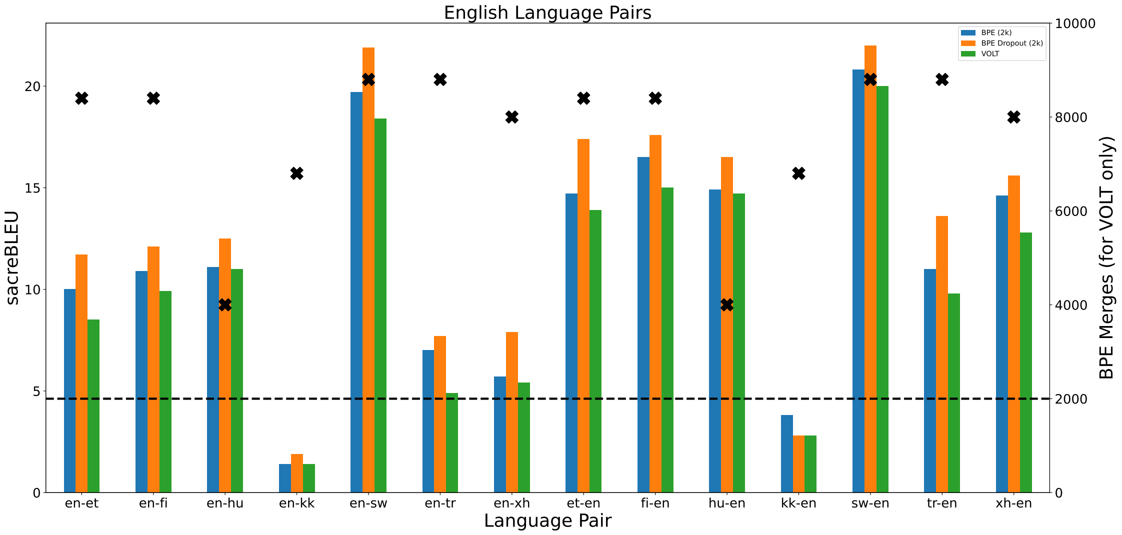

Figure 2: sacreBLEU scores for different subword segmentation strategies on the English language pairs: BPE 2k (blue), BPE dropout 2k (orange) and VOLT (green). The crosses mark the BPE merge size chosen by VOLT, for easy comparison to the baselines (2000).

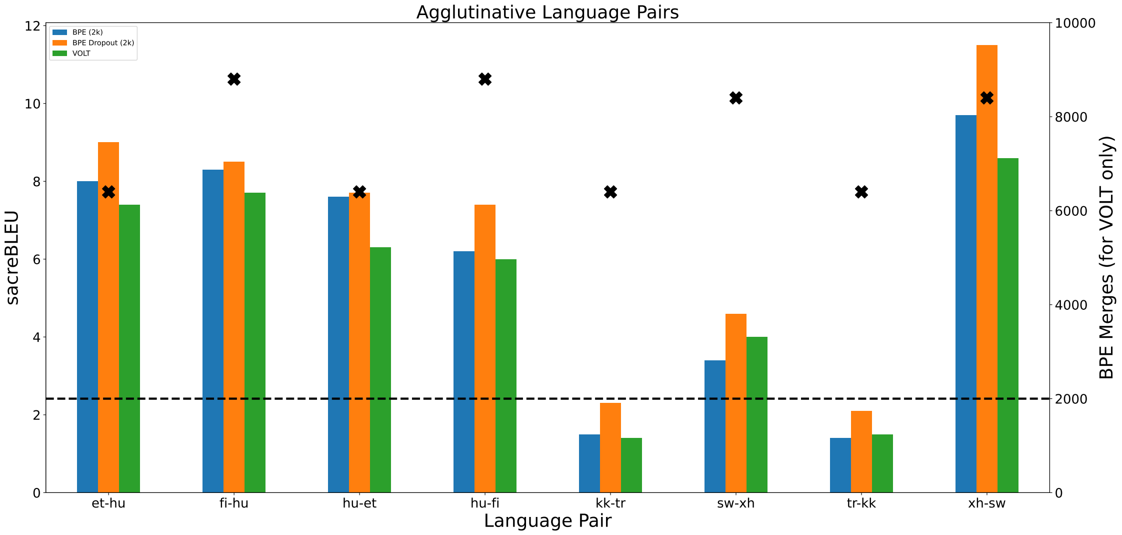

Figure 3: sacreBLEU scores for different subword segmentation strategies on the non-English agglutinative language pairs: BPE 2k (blue), BPE dropout 2k (orange) and VOLT (green). The crosses mark the BPE merge size chosen by VOLT, for easy comparison to the baselines (2000).

Results Results are presented on FLORES 'devtest' sets. We discluded our results for all language pair involving Tamil as we could not get these to train; on further inspection of the training data, we found that the JW300 corpora suffered from severe misalignment. We suspect the same has happened with Kazakh (kk) language pairs.

We split our results between English and non-English language pairs for clarity (Figures 2 and 3)

[3] `transformer_wmt_de_en`

[4] the derivation and details of which can be found in the paper

respectively. We also evaluated our models using chrF (Popovic, 2015) and the pretrained metric COMET (Rei et al., 2020), but as they correlated near perfectly with sacreBLEU (Post, 2018), we only report sacreBLEU. The figures show that BPE dropout reliably outperforms BPE, thus confirming the results of Provilkov et al. (2020) that BPE dropout does indeed improve performance in a low-resource setting. Furthermore, this pattern is consistent irrespective of whether the machine translation is agglutinative to agglutinative or not.

However, VOLT rarely outperforms the BPE baseline (and bear in mind that this BPE had been chosen without any expensive hyperparameter tuning, something Xu et al. (2021) claimed to solve the need for). The VOLT authors chose a relatively high BPE merge count (around 8000 merges) as their baseline, which we think is the reason for VOLTs apparent effectiveness in their work. We followed the findings of Ding et al. (2019), where they recommend much lower BPE merge/vocab size for low resource MT.

3.1.3 Concluding Remarks

Our key takeaway from this work is that BPE dropout performs remarkably well across a range of scenarios, and should always be the baseline for any work on low-resource MT. Despite being a best paper winner at ACL, VOLT (Xu et al., 2021) did not perform as expected, and our experiments suggest that the baselines used in the paper were weak. During our experiments we also found discrepancies between the reported model, and the implementation, which were confirmed in correspondence with the authors.

4 Factors encoding latent features of morphology

4.1 Morphological Typology and Machine Translation

In MT research, languages are often labelled with terms like “fusional”, or “agglutinative”, without giving a precise definition of those terms. Recent work on morphological typology (Payne, 2017) has argued that morphological properties of language can be measured on a continuous scale, at the segment level. We follow Payne (2017) in considering two indices: synthesis and fusion. Synthesis is a measure of the number of morphemes per word, and ranges from from analytic (low synthesis) to polysynthetic (high synthesis). Fusion measures the number of fusional morpheme joints (i.e. joints where morphemes are fused rather than concatenated) and varies from highly fusional to highly agglutinative.

Considering three different language pairs (English–Spanish, English–Turkish and English–German) we develop tools to measure synthesis and fusion and so observe their effect on MT. In order to measure synthesis, we need to use a morphological analyser, whereas we can measure fusion (for Spanish) by using a part-of-speech tagger to identify verbs and a set of hand-written rules to cover all verb paradigms.

Our analysis shows that a higher degree of synthesis or fusion usually corresponds to less accurate translations in specific word types (studying nouns and verbs in English–Turkish, and verbs in English–Spanish). At segment level, we show that synthesis and fusion-based predictors correlate with MT quality across the 3 language pairs we studied (in both directions).

This work was published at NAACL 2022 (Oncevay et al., 2022), so we refer the reader to the paper for more details.

5 Conclusion

This deliverable has described several pieces of work addressing the problem of morphology in MT. We have shown how to improve low-resource MT with word-level tags on the target, compared unsupervised word splitting algorithms in the translation of agglutinative languages, demonstrated how morphological typology can be used to analyse MT, and released a state-of-the-art lemmatisation and morphological analysis toolkit.

Papers

The following papers have resulted from the work of WP3 in the second half of the project, i.e since July 2020.

- *Quantifying Synthesis and Fusion and their Impact on Machine Translation* Oncevay et al. (2022)

Code and Data Releases

- Morphological Tagging and Lemmatization in Context - code to reproduce state-of-the-art morphological taggers and lemmatizers. https://github.com/IvoOVerhoeven/morph_tag_lemmatize

References

Agic, Z. and Vulic, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3204–3210. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ding, S., Renduchintala, A., and Duh, K. (2019). A call for prudent choice of subword merge operations in neural machine translation. In Forcada, M. L., Way, A., Haddow, B., and Sennrich, R., editors, *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*, pages 204–213. European Association for Machine Translation.

Feng, X., Feng, Z., Zhao, W., Zou, N., Qin, B., and Liu, T. (2019). Improved neural machine translation with pos-tagging through joint decoding. In Han, S., Ye, L., and Meng, W., editors, *Artificial Intelligence for Communications and Networks*, pages 159–166, Cham. Springer International Publishing.

García-Martínez, M., Barrault, L., and Bougares, F. (2016). Factored neural machine translation architectures. In *Proceedings of the 13th International Workshop on Spoken Language Translation*.

Goyal, N., Gao, C., Chaudhary, V., Chen, P., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2021). The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

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

McCarthy, A. D., Silfverberg, M., Cotterell, R., Hulden, M., and Yarowsky, D. (2018). Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101. Association for Computational Linguistics.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R., and Hulden, M. (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Nadejde, M., Reddy, S., Sennrich, R., Dwojak, T., Junczys-Dowmunt, M., Koehn, P., and Birch, A. (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. Association for Computational Linguistics.

Oncevay, A., Ataman, D., van Berkel, N., Haddow, B., Birch, A., and Bjerva, J. (2022). Quantifying Synthesis and Fusion and their Impact on Machine Translation. In *Proceedings of NAACL*.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In Ammar, W., Louis, A., and Mostafazadeh, N., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Payne, T. E. (2017). Morphological typology. In *The Cambridge Handbook of Linguistic Typology*, pages 78–94. Cambridge University Press.

Popovic, M. (2015). chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Provilkov, I., Emelianenko, D., and Voita, E. (2020). Bpe-dropout: Simple and effective subword regularization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1882–1892. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Sánchez-Cartagena, V. M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2020). Understanding the effects of word-level linguistic annotations in under-resourced neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3938–3950, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sennrich, R. and Haddow, B. (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. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Sprugnoli, R., Passarotti, M., Cecchini, F. M., and Pellegrini, M. (2020). Overview of the EvaLatin 2020 evaluation campaign. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association (ELRA).

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Straka, M., Straková, J., and Hajic, J. (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. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2021). Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Wagner, M. (2017). Target factors for neural machine translation.

Xu, J., Zhou, H., Gan, C., Zheng, Z., and Li, L. (2021). Vocabulary learning via optimal transport for neural machine translation. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7361–7373. Association for Computational Linguistics.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D2.2 GoURMET Final report on modelling morphological structure