The Lazy Encoder: A Fine-Grained Analysis of the Role of Morphology in Neural Machine Translation

Arianna Bisazza
Leiden University
The Netherlands
a.bisazza@liacs.leidenuniv.nl

Clara Tump
KTH Royal Institute of Technology
Sweden
clara.tump@hotmail.com

Abstract

Neural sequence-to-sequence models have proven very effective for machine translation, but at the expense of model interpretability. To shed more light into the role played by linguistic structure in the process of neural machine translation, we perform a fine-grained analysis of how various source-side morphological features are captured at different levels of the NMT encoder while varying the target language. Differently from previous work, we find no correlation between the accuracy of source morphology encoding and translation quality. We do find that morphological features are only captured in context and only to the extent that they are directly transferable to the target words.

1 Introduction

The advent of Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014) has led to remarkable improvements in machine translation quality (Bentivogli et al., 2016) but has also produced models that are much less interpretable. In particular, the role played by linguistic features in the process of understanding the source text and rendering it in the target language remains hard to gauge. Acquiring this knowledge is important to inform future research in NMT, especially regarding the usefulness of injecting linguistic information into the NMT model, e.g. by using supervised annotation (Sennrich and Haddow, 2016).

Hill et al. (2014) gave a first answer to this question, reporting high accuracies by source-side NMT word embeddings on the well-known analogy task by Mikolov et al. (2013) which also includes a number of derivational and inflectional transformations in the morphologically poor English language. More recent work (Shi et al., 2016) has shown that source *sentence* representations produced by NMT encoders contain a great deal of syntactic information. Belinkov et al. (2017a) focused on the *word* level and examined to what extent part-of-speech and morphological information can be extracted from various NMT word representations. The latter study found that source-side morphology is captured slightly better by the first recurrent layer than by the word embedding and the final recurrent layer. Another, somewhat surprising finding was that source-side morphology is learned better when translating into an 'easier' target language than into a related one, even if the 'easier' language is morphologically poor.

In this paper, we also focus on source-side morphology but perform a finer-grained analysis of how morphological features are captured by different components of the NMT encoder while varying the target language. We argue that predicting generic morphological tags where all features are mixed, as done by Belinkov et al. (2017a), can only give us a limited insight into the linguistic competence of the model. Hence, we predict morphological features independently from one another and ask the following questions:

- Are different morphological features captured by the NMT encoder to substantially different extents and, if yes, why?
- Are morphological features captured as a word type property (i.e. at the word embedding level) or are they mostly computed in context (i.e. at the recurrent state level)?
- How does source-target language relatedness affect the morphological competence of the NMT encoder?

More specifically, we look at whether the NMT encoder only learns those morphological features that can be directly transferred to the target words

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2871–2876
Brussels, Belgium, October 31 - November 4, 2018.

Figure 1: Two example French sentences translated to Italian, German, and English. Number (top) is usually carried over to the three target languages, while gender (bottom) is less predictable. Colored fonts mark agreement with the noun in boldface.

(such as *number*) or whether it also learns features that are not directly transferable but can still be useful to correctly parse and infer the meaning of a sentence (such as *gender*). See example in Fig. 1.

We focus on French and similarly to previous work (Shi et al., 2016; Belinkov et al., 2017a) we use the continuous word representations produced by a trained NMT system to build and evaluate a number of linguistic feature classifiers. Classifier accuracy represents the extent to which a given feature is captured by the NMT encoder.

2 Methods

We train NMT systems on the following language pairs: French-Italian (FR_{IT}), French-German (FR_{DE}), and French-English (FR_{EN}). We chose these language pairs for their different levels of language relatedness and morphological feature correspondence. Grammatical gender is especially interesting as it is marked in French, Italian and German, but not in English (except for a few pronouns). The gender of Italian nouns often corresponds to that of French because of their common language ancestor, whereas German gender is mostly unrelated from French gender (see example in Fig. 1).

The continuous word representations produced by the three NMT systems while encoding a corpus of French sentences are used to build and evaluate several specialized classifiers: one per morphological feature. If a classifier significantly outperforms the majority baseline, we conclude that the corresponding feature is captured by the NMT encoder. While this methodology is similar to that of previous work (Köhn, 2015; Belinkov et al., 2017a,b; Dalvi et al., 2017) we make sure that our results are not affected by overfitting by eliminating any vocabulary overlap between the classifier's training and test sets. We find this step crucial to ensure that the redundancy in this type of data does not lead to over-optimistic conclusions. We now provide more details on the experimental setup.

Parallel corpora. For a fair comparison among target languages, we extract the intersection of the Europarl corpus (Koehn, 2005) in our three language pairs so that the source side data is identical for all NMT systems. Sentences longer than 50 tokens are ignored. This data is then split into an NMT training, validation, and test set of 1.3M, 2.5K, and 2.5K sentence pairs respectively.

NMT model. The NMT architecture is an attentional encoder-decoder model similar to (Luong et al., 2015) and uses a long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) as the recurrent cell. The models have 3 stacked LSTM layers and are trained for 15 epochs. Embedding and hidden state sizes are set to 1000. Source and target vocabularies are limited to the 30,000 most frequent words on each side of the training data.[1] The NMT models achieve a test BLEU score of **32.6**, **25.4** and **39.4** for French-Italian, French-German and French-English respectively.

Continuous word representations. Given a source sentence, the NMT system first encodes it into a sequence of word embeddings (context-independent representations), and then into a sequence of recurrent states (context-dependent representations). As we are mostly interested in the impact of context on word representations, we compare the word embeddings against the final layer of the stacked LSTMs (corresponding to layers 0 and 3 in Belinkov et al. (2017a)'s terms) while disregarding the intermediate layers.

Morphological classification. The continuous word representations are used to train a logistic regression classifier[2] for each morphological feature: *gender* and *number* for noun and adjectives; *tense* for verbs (with labels: present, fu-

[1] Subword/character-level representations are not included in this study since we are interested in the models' ability to learn morphology from word usage, rather than word form.

[2] We use linear classifiers since their accuracies can be interpreted as a measure of supervised clustering accuracy, which gives a better insight on the structure of the vector space (Köhn, 2015). Results with a simple multi-layer perceptron were consistent with the findings by the linear classifier, with slightly better performance overall.

ture, imperfect, or simple past). Word labels are taken from the *Lefff* French morphological lexicon (Sagot, 2010)[3]. To ensure a fair comparison between context-independent and context-dependent embedding classification, words that are ambiguous with respect to a given feature are excluded from the respective classifier's training and test data.

Classifiers' training/test data. The classifiers are trained on a 50K-sentence subset of the NMT training data and tested on the NMT test sets (2.5K). For each experiment, we extract *one vector per token* from the NMT encoder. While this is the only possible setup for context-dependent representations, it leads to a problematic training/test overlap in the word embedding experiment because all occurrences of the same word are associated to exactly the same vector. We find that, due to this overlap, a dummy binary feature assigned to a random half of the vocabulary can be predicted from the word embeddings with very high accuracy (86% for a linear, 98% for a non-linear classifier) leading to over-optimistic conclusions on the linguistic regularities of these representations. To avoid this, we split the vocabulary in two parts of 15K types: the first is used to filter the training samples and the second to filter the test samples. We repeat each experiment five times using five different random vocabulary splits and report mean accuracies. This process is applied to all experiments (including those on hidden states) to allow for a fair comparison of the results.

3 Results and Discussion

This section presents our results along three dimensions: context-dependency of the word representations (§3.1), different morphological features (§3.2), and target language impact (§3.3). Unless explicitly stated, all discussed results are statistically significant (computed using a t-test for a one-tailed hypothesis and independent means).

3.1 Word embeddings *vs* recurrent states

One of our goals was to discover whether morphological features are captured as a word type property or in context. Fig. 2 shows the extent to which the NMT encoder captures different features at the word level (word embeddings) compared to the recurrent state level (LSTM state), averaged over all target languages. We can see that each feature is clearly captured at the recurrent state level, confirming that source-side morphology is indeed successfully exploited by NMT models. However, at the word embedding level, accuracies are comparable to the majority class baseline (these differences are not significant), which implies that the source-side lexicon of our NMT systems does not encode morphology in a systematic way. This might be partly explained by the fact that learning morphological features at the word level is difficult due to data sparsity – indeed the rarest French words in our dataset are observed only 10 times in the training data. However, additional experiments showed that our finding is consistent across different word frequency bins: that is, even the embeddings of frequent words do not encode morphological features better than the majority baseline.

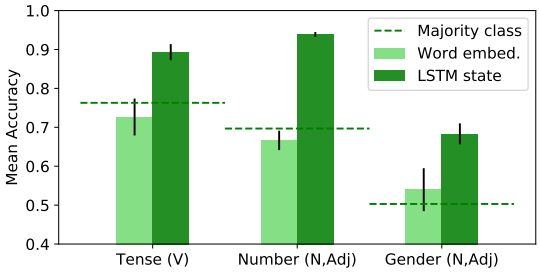

Figure 2: Classifier accuracy for different morphological features, averaged over target languages.

This result is surprising, considering that our morphological features are usually easy to infer from the immediate context of French words (see examples in Fig.1) and that morphology was shown to be well captured by monolingual word embeddings in various European languages including French (Köhn, 2015). By contrast, our NMT encoders choose not to store morphology at the word type level, perhaps in order to allocate more capacity to semantic information.

3.2 Different morphological features

Secondly, we asked whether the NMT encoder captured different morphological features to different extents. For this question, we disregard the word embedding results because none of the features are significantly captured at this level.

Fig. 2 shows that the mean accuracy of *number* is the highest, followed by *tense* and then by *gender*. However, it should be noted that the majority baselines for number and tense are much higher than the one for gender. In both absolute and rela-

[3]Lexique des Formes Fléchies du Français: http://alpage.inria.fr/~sagot/lefff-en.html

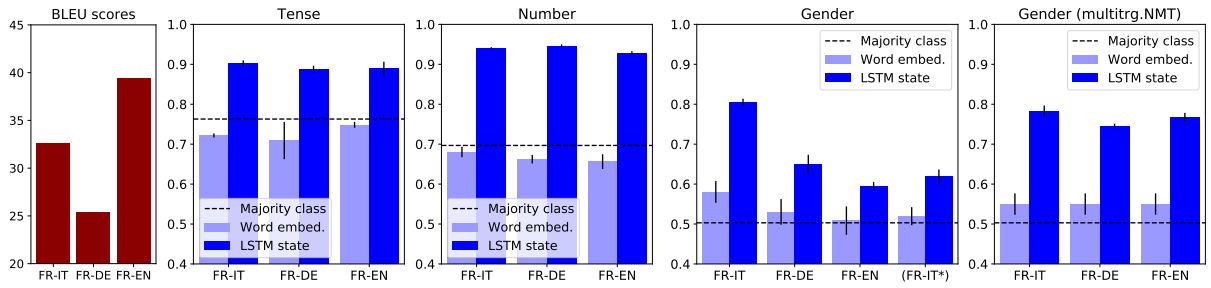

Figure 3: Translation quality (left) and classifier accuracy for each morphological feature in the three different language pairs. IT* denotes a modified Italian language where gender marking is removed.

tive terms, the best performing feature is number. This can be explained by the fact that number remains most often unchanged through translation, and is marked in all target languages – albeit to different extents. On the other hand, tense is determined by the semantics but also by language-specific usage, while gender has little semantic value and is mostly assigned to nouns arbitrarily.

The fact that the results of different morphological features are so variable confirms the setup of examining each feature independently.

3.3 Source-target language relatedness

Fig. 3 shows the impact of the target language on the encoded morphology accuracy. We again focus our analysis on the LSTM state level since embedding level results are mostly near the baseline.

Differently from Belinkov et al. (2017a) we do not find that source-side morphology is captured better when translating into the 'easiest' language, which in our case is English, both in terms of morphological complexity and BLEU performance. We note that their findings were based on very small, possibly not significant differences, and on the prediction of all morphological features simultaneously. By contrast, our fine-grained analysis reveals that the impact of target language is significant and even major on only one feature, namely *gender*, where it agrees with our linguistic intuition. Indeed this feature differs from the others because it varies largely among languages and, when present, is semantically determined only to a very limited extent. FR_{IT}, where source gender is a good predictor of target gender, shows the highest accuracy; FR_{EN}, where target gender is not marked, shows the lowest; FR_{DE}, where source gender is often confounding for target gender, lies in-between.

Is language relatedness the main explaining variable? To find that out, we experiment with a modified Italian target language without gender marking, i.e. all gender-marked words are replaced by their masculine form (FR_{IT^*}). This language pair achieves a slightly higher BLEU score than FR_{IT} (33.2 *vs* 32.6), which can be attributed to the smaller target vocabulary. However its source gender accuracy is much worse (see Fig. 3), which indicates that the high performance of the FR_{IT} encoder is mostly due to the ubiquitous gender marking in the target language, rather than to language relatedness. All this suggests that source morphological features contribute to sentence understanding to some degree, but the incentive to learn them mostly depends on how directly they can be transferred to the target sentence.

Finally, we look at what happens when a single NMT system is trained in a multitarget fashion on our three language pairs. Following the setup of Johnson et al. (2017), we prepend a *to-target-language* tag $\{2it, 2de, 2en\}$ to the source side of each sentence pair and mix all language pairs in the NMT training data. Results are presented for gender in Fig. 3 (right).[4] Note that, while word embeddings are identical for the three language pairs, recurrent states change according to the language tag. In this setup the target language impact is less visible and gender accuracy at the LSTM state level is overall much higher than that of the mono-target systems (0.77 *vs* 0.68 on average) whereas BLEU scores are slightly lower (-0.9% on average). While this is only an initial exploration of multilingual NMT systems, our results suggest that this kind of multi-task objective pushes the model to learn linguistic features in a more consistent way (Bjerva, 2017; Enguehard et al., 2017).

[4] Multitarget results for tense and number did not differ significantly from the corresponding monotarget results.

4 Conclusion

We have confirmed previous findings that morphological features are significantly captured by word-level NMT encoders. However, the features are not captured at the word type level but only at the recurrent state level where word representations are context-dependent. Secondly, there is a visible difference in the extent to which different morphological features are learned: Semantic categories like number and verb tense are well captured in all language pairs, whereas grammatical gender with its only agreement-triggering function, is dramatically affected by the target language. Source-side gender is encoded well only when it is a good predictor of target gender and when target-side marking is extensive, i.e. when translating from French to Italian.

Our findings indicate that the importance of linguistic structure for the neural translation process is very variable and language-dependent. They also suggest that the NMT encoder is rather 'lazy' when it comes to learning grammatical features of the source words, unless these are directly transferable to their target equivalents.

Acknowledgments

This research was partly funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646. Part of the work was carried out on the DAS computing system (Bal et al., 2016) while the authors were affiliated at the Informatics Institute of the University of Amsterdam. We thank Ke Tran for providing feedback on the early stages of this research.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Henri Bal, Dick Epema, Cees de Laat, Rob van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, and Harry Wijshoff. 2016. A medium-scale distributed system for computer science research: Infrastructure for the long term. *Computer*, 49(5):54–63.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas. Association for Computational Linguistics.

Johannes Bjerva. 2017. *One Model to Rule them All: Multitask and Multilingual Modelling for Lexical Analysis*. Ph.D. thesis, University of Groningen.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, and Stephan Vogel. 2017. Understanding and improving morphological learning in the neural machine translation decoder. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 142–151. Asian Federation of Natural Language Processing.

Émile Enguehard, Yoav Goldberg, and Tal Linzen. 2017. Exploring the syntactic abilities of rnns with multi-task learning. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 3–14, Vancouver, Canada. Association for Computational Linguistics.

Felix Hill, KyungHyun Cho, Sebastien Jean, Coline Devin, and Yoshua Bengio. 2014. Not all neural embeddings are born equal. In *NIPS 2014 Workshop on Learning Semantics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernand a Vigas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Arne Köhn. 2015. What's in an embedding? analyzing word embeddings through multilingual evaluation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages

2067–2073, Lisbon, Portugal. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Benoît Sagot. 2010. The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *7th international conference on Language Resources and Evaluation (LREC 2010)*.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1526–1534, Austin, Texas. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.