Contextualization of Morphological Inflection

Ekaterina Vylomova[ϑ] **Ryan Cotterell**[?,fi] **Timothy Baldwin**[ϑ] **Trevor Cohn**[ϑ] and **Jason Eisner**[?]

[ϑ]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[fi]The Computer Laboratory, University of Cambridge, Cambridge, UK
[?]Department of Computer Science, Johns Hopkins University, Baltimore, USA
{vylomovae,tbaldwin,tcohn}@unimelb.edu.au
rdc42@cam.ac.uk jason@cs.jhu.edu

1 Introduction

NLP systems are often required to generate grammatical text, e.g., in machine translation, dialogue, and grammar correction. One component of grammaticality is the use of contextually appropriate closed-class morphemes. We study **contextual inflection**, which has been recently introduced in the CoNLL-SIGMORPHON 2018 shared task (Cotterell et al., 2018) to directly investigate context-dependent morphology in NLP. There, a system must inflect lemmatized tokens in sentential context. For example, in English, the system must reconstruct the correct word sequence *two cats are sitting* from partially lemmatized sequence *two _cat_ are sitting*. Among other things, this requires: (1) identifying *cat* as a noun in this context, (2) recognizing that *cat* should be inflected as plural to agree with the nearby verb and numeral, and (3) realizing this inflection as the suffix *s*.

The task focuses on just the ability to reconstruct certain missing parts of the sentence—inflectional morphemes and their orthographic realization. Contextual inflection does not perfectly separate grammaticality modeling from content modeling: mapping *two cats _be_ sitting* to the fully-inflected *two cats were sitting* does not require full knowledge of English grammar—the system does not have to predict the required word order nor the required auxiliary verb *be*, as these are supplied in the input. Conversely, this example does still require predicting some content—the semantic choice of past tense is *not* given by the input and must be guessed by the system.This morphological feature is *inherent* in the sense of Booij (1996).

Here we evince a simple point: models are better off jointly predicting morphological tags from context than directly learning to inflect lemmata from sentential context. We provide an analysis discussing the role of morphological complexity in model performance. We additionally note that most uncertainty comes from inherent morphological categories.

2 Experiments

Dataset. We use the Universal Dependencies v1.2 dataset (Nivre et al., 2016) for our experiments. We include all the languages with information on their lemmata and fine-grained grammar tag annotation that also have `fasttext` embeddings (Bojanowski et al., 2017), which are used for word embedding initialization.

Models and Evaluation. We compare the following models: (1) a novel structured neural model for contextual inflection ("JOINT"). The model predicts the sequence of morphological tags from the lemmatized sequence (using neural parameterization of CRF similar to Lample et al. (2016)) and then inflects the target lemmata using the reinflection model from Aharoni and Goldberg (2017); (2) a neural encoder–decoder with an attention mechanism ("SM"; Cotterell et al. (2018)), where the encoder represents a target form context as a concatenation of its lemma, its left and right word forms, their lemmata and tag representations, and then the decoder generates the target inflected form character-by-character; and (3) a monolingual version of the best performing system of the shared task ("CPH"; Kementchedjhieva et al. (2018)) that augments the above encoder–decoder with sentence-level left and right contexts (comprising of forms, their lemmata and morphological tags) as well as predicts morphological tags for a target form as an auxiliary task. We additionally evaluate SM model on prediction of the target form without any information on morphological tags ("DIRECT"). [1]

[1]For "CPH" and "SM" the hyperparameters are set as described in Cotterell et al. (2018). The joint and "DIRECT"

Language	tag	form				
	JOINT	GOLD	JOINT	DIRECT	SM	CPH
Bulgarian	81.6	91.9	**78.8**	71.5	77.1	76.9
English	89.6	95.6	**90.4**	86.8	86.5	86.7
Basque	66.6	82.2	61.1	59.7	**61.2**	60.2
Finnish	66.0	86.5	**59.3**	51.2	56.6	56.4
Gaelic	68.3	84.5	**69.5**	64.5	68.9	66.9
Hindi	85.3	88.3	81.4	85.4	86.8	**87.5**
Italian	92.3	85.1	80.4	85.2	88.7	**90.5**
Latin	82.6	89.7	**75.7**	71.4	74.2	74.9
Polish	71.9	96.1	**74.8**	61.8	72.4	70.2
Swedish	81.9	96.0	**82.5**	75.4	78.4	80.9

Table 1: Accuracy of the models for various prediction settings. **tag** refers to tag prediction accuracy, and **form** to form prediction accuracy. GOLD denotes form prediction conditioned on gold target morphological tags.

We evaluate models' ability to predict: (i) correct morphological tags, and (ii) correct inflected forms. As our evaluation metric, we report 1-best accuracy for both tags and word form prediction.

3 Results and Discussion

Tab. 1 presents the accuracy of our best model across all languages. Below we highlight two main lessons from our error analysis that apply to a wider range of generation tasks, e.g., machine translation and dialog systems.

Directly Predicting Morphology. Tab. 1 indicates that all systems that make use of morphological tags outperform the DIRECT baseline on most languages. The comparison of the joint model with latent morphological tags to the direct form generation in SM suggests that we should be including linguistically-motivated latent variables into models of natural language generation. We observe in Tab. 1 that predicting the tag together with the form often improves performance.

Morphological Complexity Matters. We observed that for languages with rich case systems, e.g., the Slavic languages (which exhibit a lot of fusion), the agglutinative Finno-Ugric languages, and Basque, performance is much worse. These languages present a broader decision space and often require inferring which morphological categories need to be in agreement in order to make an accurate prediction. This suggests that generation in languages with more morphological complexity will be a harder problem for neural models to solve.

Error Analysis. Our analysis of adjective–noun agreement category prediction suggests that JOINT model is able to infer adjective number, gender, and case from its head noun. Verb gender, which appears only in the past tense of many Slavic languages, seems to be harder to predict. Given that the linear distance between the subject and the verb may be longer, we suspect the network struggles to learn longer-distance dependencies, consistent with the findings of Linzen et al. (2016). Overall, automatic inference of agreement categories is an interesting problem that deserves more attention, and we leave it for future work.

We also observed that most uncertainty comes from morphological categories such as noun number, definiteness,[2] and verb tense, all of which are inherent (Booij, 1996)[3] and typically cannot be predicted from sentential context if they do not participate in agreement.[4] On the other hand, aspect, although being closely related to tense, is well-predicted since it is mainly expressed as a separate lexeme.[5]

In addition, we evaluated the joint model's performance when *all* forms are replaced by their corresponding lemmata (as in *two cat be sit*). For freer word order languages such as Polish or Latin, we observe a substantial drop in performance because most information on inter-word relations and their roles (expressed by means of case system) is lost.

4 Conclusion

Our analysis demonstrated that the contextual inflection can be a highly challenging task, and the inclusion of morphological features prediction is an important element in such a system. We also highlighted two types of morphological categories, contextual and inherent, in which the former relies on agreement and the latter comes from a speaker's intention.

models use word and character embedding dimensionalities of 300 and 100, respectively. The hidden state dimensionality was set to 200. All models were trained with Adam (Kingma and Ba, 2014) for 20 epochs.

[2] which is expressed morphologically in Bulgarian

[3] Such categories exist in most languages that exhibit some degree of morphological complexity.

[4] Unless there is no strong signal within a sentence such as *yesterday*, *tomorrow*, or *ago* as in the case of tense.

[5] But, in general, it is still problematic to make a prediction in languages where aspect is morphologically marked or highly mixed with tense as in Basque.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2004–2015.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Geert Booij. 1996. Inherent versus contextual inflection and the split morphology hypothesis authors. In *Yearbook of Morphology 1995*, pages 1–16. Springer, Netherlands.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Gèraldine Walther, Ekaterina Vylomova, Arya D McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 Shared Task: Universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27.

Yova Kementchedjhieva, Johannes Bjerva, and Isabelle Augenstein. 2018. Copenhagen at CoNLL–SIGMORPHON 2018: Multilingual inflection in context with explicit morphosyntactic decoding. *CoNLL–SIGMORPHON*, page 93.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC 2016*.