

KU-CST at CoNLL–SIGMORPHON 2018 Shared Task: a *Tridirectional* Model

Manex Agirrezabal

Centre for Language Technology (CST)
Department of Nordic Studies and Linguistics (NorS)
University of Copenhagen (KU)
manex.agirrezabal@hum.ku.dk

Abstract

In this paper we describe our sequence-to-sequence model for morphological inflection. We have constructed a common Encoder-Decoder network that encodes the input lemma into a dense vector to translate it to an inflected form, based on input morphological tags. The main novelty of the model is that the input lemma is encoded in three different directions: left-to-right, right-to-left and boundaries-to-center. In this paper we report the accuracies of the model compared to the same bidirectional approach.

1 Introduction

In this work we present the neural network architecture prepared for the task of morphological inflection in the CoNLL–SIGMORPHON 2018 Shared Task (Cotterell et al., 2018). Both morphological analysis and morphological inflection are crucial in end-to-end Natural Language Processing pipelines, as they are one of the initial steps performed before solving more high-level problems such as Named-Entity Recognition, Sentiment Analysis, or others.

2 Task

In the CoNLL–SIGMORPHON 2018 Shared Task there were two tasks to solve. In this work we present a possible solution for the first task, in which word-forms have to be built without considering the context. The input in the task is a lemma and a list morphological tags. The system should be then able to produce the corresponding word form. The following is an example from the Spanish dataset

jaquear
V; COND; 1; PL
↓
jaqueamos

in which the input lemma is *jaquear* and its morphological tags state that the word form should be a verb (V) in conditional tense (COND), first person (I) in plural (PL). The output word form is *jaqueamos*.

The models can be trained and tested in over 100 languages and in three different settings, low-, medium- and high-resource scenarios (100, 1,000 or 10,000 training instances, respectively).

2.1 Dataset

As mentioned above we trained and tested our models in the provided dataset, which contains morphological inflections for over 100 languages. The information is encoded using Unicode and morphological tags follow the UniMorph tagging schema (Kirov et al., 2018).

3 Method

Following previous successful attempts to morphological inflection (Kann and Schütze, 2016), we built a model based on Neural Networks, specifically an Encoder-Decoder network (Cho et al., 2014) with an attention mechanism (Bahdanau et al., 2014). Furthermore, instead of constructing a linguistically inspired model, we have shortly explored an engineering approach. The main novelty of our model is in the way the input is encoded.

In Lample et al. (2016) it is stated that recurrent architectures such as Recurrent Neural Networks are capable of encoding very long sequences, but the representation is biased towards the last explored items. Because of that, a bidirectional RNN could be expected to represent well the structure of a word, as it models both the ending (suffix) and the beginning (prefix) by the use of a forward encoder and a backward encoder, respectively.

Our model explores whether this architecture can be improved adding another encoder that en-

codes the word starting at the boundaries and ending in the center, capturing in that way the central structure of the word. The architecture is shown in Figure 1.

Each input is encoded with three different encoders, left-to-right, right-to-left and boundaries-to-center, and those encoded representations are concatenated. Then, a many-hot encoding representation of the morphological tags is concatenated at the end. In this way, we generate the representation of our source lemma with its morphological information. This representation goes to the decoder so that the output word is generated character by character.

4 Model configuration

The implementation is based on a Machine Translation model created using the Pytorch framework. It encodes sentences using three Recurrent Neural Networks with 128 GRU cells in each encoder. In order to train the decoder, we use a teacher forcing ratio of 0.5. We started training all the models for 10 epochs, but we could observe that the models from the low-resource scenario did not converge and the ones in the high-resource scenario did not improve results after the fifth epoch. Because of that, we train our models for 20, 15 and 5 epochs in the low-, medium- and high-resource scenarios, respectively.

5 Results

We tested a bidirectional and a *tridirectional* model and our expectations were that the tridirectional one would show a better performance in the development set. As you can see in Table 1, the mean accuracy is slightly better than in the bidirectional model with the same exact configuration,¹ although these differences are not significant according to a bootstrap test.

In Table 2 you can see the accuracies for each language in each setting with the tridirectional approach². In Figure 2 we plotted these accuracies together with the accuracies of the bidirectional model. In some cases, our tridirectional approach is sufficiently more accurate than the bidirectional

¹Same number of epochs, same cell types, and same size of hidden memory size. We are aware, although, that the tridirectional approach has more parameters because it has three encoders instead of two.

²In order to make it more interpretable, we marked in bold results that are better than the current baseline presented for the shared task.

Mean accuracies			
	low	medium	high
Bidirectional model	2.3	28.4	54.8
Tridirectional model	2.8	32.4	55.6
Mean Levenshtein distances			
	low	medium	high
Bidirectional model	5.25	2.30	1.28
Tridirectional model	5.21	2.11	1.41

Table 1: Average results for the bidirectional and tridirectional approaches in three different settings.

one, such as in *Crimean-Tatar*, *Galician*, *Quechua* and *welsh* in the high setting, or *Bashkir*, *Crimean-Tatar*, *Ingrian*, *Napolitan*, *Tatar* and *Urdu* in the medium setting. There are other cases, although, that the bidirectional approach shows a higher accuracy. Check, for instance, *Persian* and *Tatar* in the high setting or *Kabardian* in the medium setting.

	low	medium	high
Adyghe	0.3	65.2	99.0
Albanian	0.2	11.9	13.4
Arabic	0.0	7.1	41.9
Armenian	0.0	13.0	48.0
Asturian	1.0	48.3	87.1
Azeri	1.0	11.0	44.0
Bashkir	1.0	68.3	62.9
Basque	0.3	44.6	69.9
Belarusian	0.5	13.0	65.5
Bengali	1.0	75.0	78.0
Breton	5.0	68.0	60.0
Bulgarian	0.0	33.2	58.1
Catalan	0.2	24.1	85.8
Classical-Syriac	0.0	67.0	76.0
Cornish	18.0	46.0	-
Crimean-Tatar	0.0	54.0	99.0
Czech	0.2	7.8	34.1
Danish	0.2	32.7	64.8
Estonian	0.1	7.1	43.3
Faroese	0.5	9.0	25.7
Finnish	0.0	0.0	1.5
Friulian	1.0	73.0	98.0
Galician	0.6	40.4	92.0
Georgian	0.2	36.2	82.2
Greek	0.2	4.2	21.2

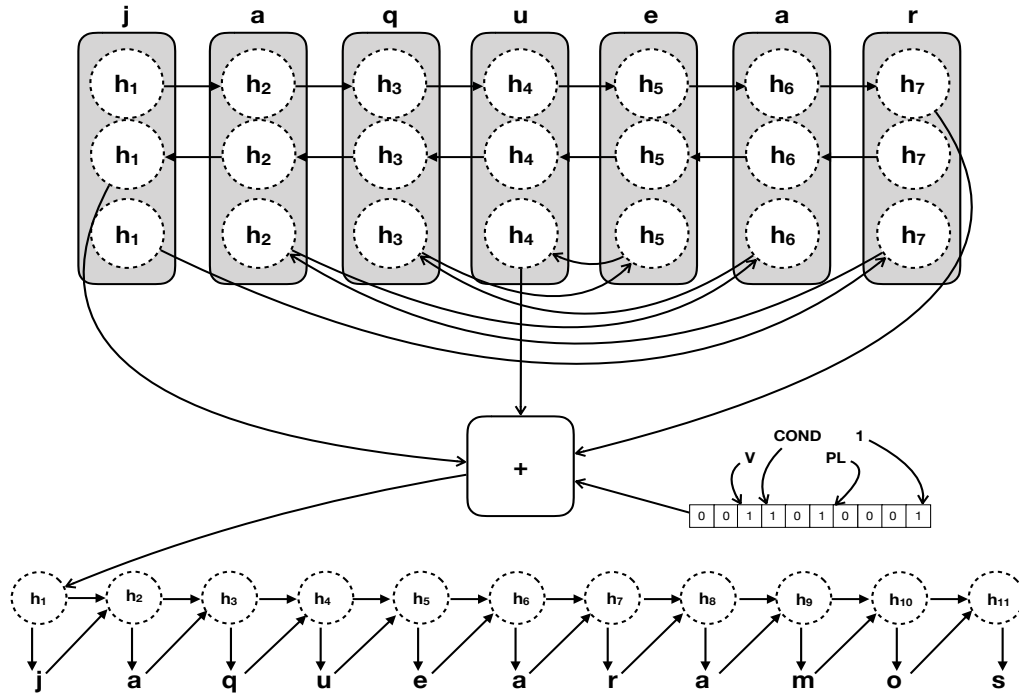


Figure 1: Architecture of our Tridirectional Encoder-Decoder model.

	low	medium	high		low	medium	high
Greenlandic	4.0	22.0	-	Middle-High-German	12.0	90.0	-
Haida	0.0	14.0	18.0	Murrinhpatha	14.0	76.0	-
Hebrew	0.4	24.3	77.9	Navajo	0.4	9.7	28.1
Hindi	1.1	73.2	84.6	Neapolitan	7.0	70.0	56.0
Hungarian	0.1	1.0	12.9	Norman	6.0	12.0	-
Icelandic	0.4	11.1	41.4	Northern-Sami	0.1	9.4	46.0
Ingrian	4.0	58.0	-	Norwegian-Bokmaal	0.8	26.2	47.8
Irish	0.2	3.6	13.8	Norwegian-Bynorsk	0.4	17.6	37.7
Italian	0.0	12.8	34.9	Occitan	1.0	40.0	99.0
Kabardian	6.0	58.0	96.0	Old-Armenian	0.0	11.8	60.7
Karelian	2.0	40.0	-	Old-Church-Slavonic	7.0	33.0	29.0
Kashubian	8.0	20.0	-	Old-French	0.0	18.2	53.9
Kazakh	0.0	22.0	-	Old-Irish	2.0	6.0	-
Khakas	2.0	22.0	-	Old-Saxon	0.6	15.1	77.1
Khaling	1.0	28.8	88.7	Pashto	2.0	48.0	85.0
Kurmanji	0.0	30.6	68.1	Persian	0.2	47.3	29.3
Ladin	5.0	60.0	93.0	Portuguese	0.2	25.9	72.6
Latin	0.0	3.5	9.6	Quechua	0.2	18.0	43.0
Latvian	0.2	8.3	50.9	Romanian	0.2	13.2	41.1
Lithuanian	0.2	3.9	37.3	Sanskrit	0.9	14.4	52.7
Livonian	1.0	25.0	57.0	Scottish-Gaelic	16.0	54.0	-
Lower-Sorbian	0.8	13.7	72.6	Serbo-Croatian	0.1	5.4	26.3
Macedonian	0.0	14.8	52.8	Slovak	0.7	10.6	59.3
Maltese	4.0	57.0	47.0	Slovene	1.2	21.2	59.3
Mapudungun	30.0	96.0	-	Sorani	0.3	28.0	29.3
Middle-French	1.1	52.8	83.1	Spanish	0.2	26.8	58.5

	low	medium	high
Swahili	4.0	70.0	90.0
Swedish	0.3	11.2	31.8
Tatar	1.0	53.0	44.0
Telugu	38.0	-	-
Tibetan	24.0	24.0	-
Turkish	0.0	7.9	23.4
Turkmen	6.0	36.0	-
Ukrainian	0.7	12.3	32.6
Urdu	4.4	79.8	95.9
Uzbek	2.0	16.0	8.0
Venetian	2.3	46.9	88.9
Votic	1.0	55.0	23.0
Welsh	0.0	23.0	91.0
West-Frisian	3.0	88.0	13.0
Yiddish	0.0	60.0	82.0
Zulu	0.3	16.2	72.7
Mean	2.8	32.4	55.6

Table 2: Accuracies for all languages in the low-, medium- and high-resource scenario using the *tridirectional* Encoder-Decoder model. The last row shows the average accuracy for each resource scenario.

6 Conclusion and Future work

In this experiment we tried to approach morphological inflection using a slightly more complex Encoder-Decoder architecture, by encoding the lemmas in three different directions (left-to-right, right-to-left and boundaries-to-center).

Although the model works quite well in some cases, there is plenty room for improvement. The main improvement that must be done is to continue experimenting with more parameters to check whether the addition of parameters improves results. Both in the medium- and high-resource settings, there are some languages that show very bad performance, especially *Finnish*, *Hungarian* and *Latin*. We feel that there is a need of carefully analyzing their outputs so that to better understand the motivation for these low results.

Data augmentation techniques were successfully used in the last CoNLL–SIGMORPHON 2018 Shared Task (Bergmanis et al., 2017; Kann and Schütze, 2017; Nicolai et al., 2017; Silfverberg et al., 2017) and thus, we think that our model could see its results improved in the low- and the medium-resource setting by adding artificially generated data.

We expect that using external resources, such

as Wikipedia, would have a positive effect. We could even analyze how much effect does a specific amount of text have in this task pretraining character embeddings with, for instance, 10,000, 50,000 or 100,000 characters.

Acknowledgments

I acknowledge the anonymous reviewers for their informative comments and possible extra future directions, and also the organizers of the CoNLL–SIGMORPHON 2018 Shared Task for giving us the opportunity of participating in such an exciting task.

References

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine 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. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, Brussels, Belgium. Association for Computational Linguistics.
- Katharina Kann and Hinrich Schütze. 2016. Med: The lmu system for the sigmorphon 2016 shared task on morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 62–70.
- Katharina Kann and Hinrich Schütze. 2017. The lmu system for the conll-sigmorphon 2017 shared task on universal morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 40–48.

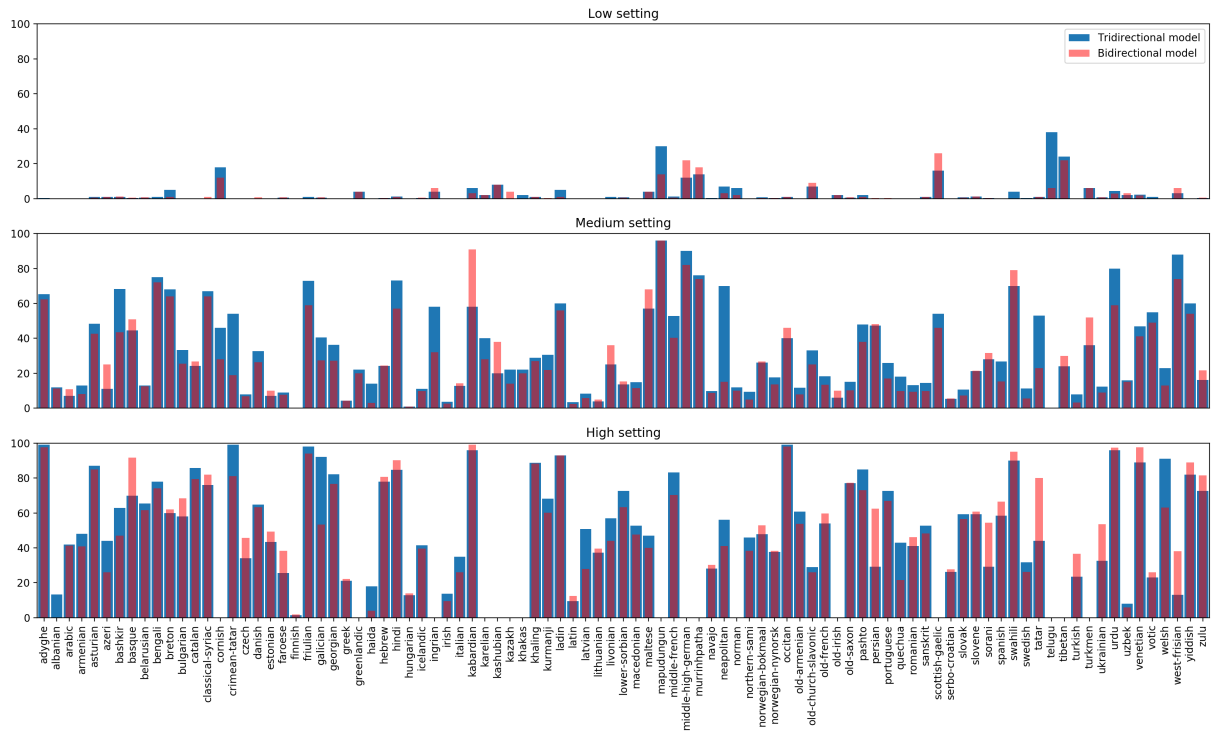


Figure 2: Accuracies of both the bidirectional and tridirectional models. As it can be seen in the figure, results from the two models are not very different as it was initially expected.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Galdine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D. McCarthy, Sandra Kbler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Garrett Nicolai, Bradley Hauer, Mohammad Motallebi, Saeed Najafi, and Grzegorz Kondrak. 2017. If you can’t beat them, join them: the university of alberta system description. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 79–84.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99.