The NYU System for the CoNLL–SIGMORPHON 2018 Shared Task on Universal Morphological Reinflection

Katharina Kann[1]
kann@nyu.edu

Stanislas Lauly[1]
stanislas.lauly@nyu.edu

Kyunghyun Cho[1,2]
kyunghyun.cho@nyu.edu

[1]Center for Data Science
New York University
New York, USA

[2]Dept. of Computer Science
New York University
New York, USA

Abstract

This paper describes the NYU submission to the CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. Our system participates in the low-resource setting of Task 2, track 2, i.e., it predicts morphologically inflected forms in context: given a lemma and a context sentence, it produces a form of the lemma which might be used at an indicated position in the sentence. It is based on the standard attention-based LSTM encoder-decoder model, but makes use of multiple encoders to process all parts of the context as well as the lemma. In the official shared task evaluation, our system obtains the second best results out of 5 submissions for the competition it entered and strongly outperforms the official baseline.

1 Introduction

The extreme type sparsity in text in a morphologically rich language, i.e., a language which relies strongly on changes in the surface form of words to express properties like gender, tense or number, requires natural language processing (NLP) systems which are able to handle inflected words in a systematic way. The SIGMORPHON and CoNLL–SIGMORPHON shared tasks on morphological reinflection, which have been held since 2016 (Cotterell et al., 2016, 2017a), encourage the development of computational models for inflection in a large number of languages.

This year's edition (Cotterell et al., 2018) features two different tasks. The datasets for Task 1 consist of triplets of lemma, morphological tag (also called the "target tag") and the corresponding inflected form, which is given for training and should be produced at test time. This is the standard inflection setup which has also been subject of the shared tasks in the last years. Task 2, in contrast, is again split into two different subtasks (called "tracks"). Both are focused on inflection in context. Here, a sentence is given, in the context of which the inflected form of which only the lemma is known should be used. The setup of the first subtask assumes that the lemmas and tags of all surrounding words are available and can be used for predicting. These might be used as desired, e.g., the tags of the previous and next words are often strong indicators for the tag of the form to be produced, which is unknown. Track 2, on the other hand, requires systems to produce inflected forms only from their lemma and the inflected context words; no tags or lemmas are given for the context. Thus, track 2 is both a more realistic and a harder version of track 1. All tasks and tracks feature 3 different settings: a low-resource setting (LOW), a medium-resource setting (MEDIUM) and a high-resource setting (HIGH).

In this paper, we describe the New York University (NYU) submission to the CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. The system we submitted was exclusively designed for Task 2, track 2, LOW. Thus, we only focus on this particular competition and do not report numbers for other setups (though, in theory, every system which works for track 2 of Task 2 can also produce output for track 1; the same holds true for LOW/MEDIUM/HIGH). Overall, our system obtains the second highest test accuracy out of 5 submitted systems and outperforms the official shared task baseline by a wide margin.

2 Morphological Inflection in Context

The system presented in this paper is designed for morphological inflection in context, i.e., predicting an inflected form which fits an indicated position in a sentence, given its lemma. Here, we will describe the task in a more formal way.

Let $\mathcal{T}$ be the set of morphological tags being

Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, pages 58–63,
Brussels, Belgium, October 31, 2018.

expressed in a language and w a lemma in the same language. We then define the morphological paradigm π of w as follows:

$$\pi(w) = \Big\{\big(f_k[w], t_k\big)\Big\}_{k \in \mathcal{T}(w)} \quad (1)$$

Here, $f_k[w]$ denotes the inflected form which corresponds to tag t_k, and both w and $f_k[w]$ are strings consisting of letters from an alphabet Σ. Note that, even though we follow the convention to describe word forms as functions of the lemma, in the huge majority of the cases, each inflected form is uniquely defined given any other word form of the same paradigm together with its morphological tag.

The task of *morphological inflection* consists of predicting a target form $f_i[w]$ from a paradigm, given the lemma w, as well as the tag t_i of the target form.

Building on this, the task of *morphological inflection in context* consists of predicting a target form $f_i[w]$ from the lemma w, as well as the context c, i.e., the sentence surrounding the target form. For the track of the shared task we are interested in, the context consists of inflected forms. Further, this task is ambiguous: for many languages, usually several morphological tags and, thus, inflected forms are acceptable for any given context.

3 Model Description

Our model is based on the standard LSTM encoder-decoder model with an attention mechanism (Bahdanau et al., 2015). Following several previous approaches (cf. Section 5), we apply it at the character level, i.e., the input to the system is the character sequence of the input lemma, represented by embeddings. The output is the (predicted) character sequence of the inflected form.

Additionally, we include the sentence context as follows: Given a sentence $s = [w_1, w_2, \ldots, w_{i-1}, l, w_{i+1}, \ldots, w_n]$, where l is the lemma of the inflected form of interest, and $w_1, \ldots, w_n$ with $n \neq i$ are the surrounding context words, we split the past context $c_{prev} = [w_1, w_2, \ldots, w_{i-1}]$ and the future context $c_{fut} = [w_{i+1}, \ldots, w_n]$ into subword units using byte pair encoding (BPE, Sennrich et al. (2016)). We then use two additional encoders to encode the sequences of subword units of both contexts.

Using bidirectional encoders, the final hidden states produced by each encoder are concatenations of the respective forward and backward hidden states:

$$h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}] \quad (2)$$

with

$$\overrightarrow{h_i} = \text{LSTM}(emb_1, \ldots, emb_i), \text{and} \quad (3)$$

$$\overleftarrow{h_i} = \text{LSTM}(emb_z, \ldots, emb_i) \quad (4)$$

$emb = emb_1, \ldots, emb_z$ represents the respective sequence of embeddings, i.e., either the embeddings of the lemma's characters or the embeddings of the subword units of either context.

Our model then uses 3 attention mechanisms—one for each encoder—to produce a context vector for each output position: H_t for the lemma, H_t^p for the past context and H_t^f for the future context.

The input to the decoder LSTM at each timestep is the concatenation of all contexts and the embedding of the last output character. Embeddings are shared between the character encoder and the decoder, BPE embeddings are shared between the two context encoders.

An overview of our model architecture is shown in Figure 1. Our final system in an ensemble of 5 random restarts of the model, combined via majority voting.

3.1 Training and Hyperparameters

Using the shared task development sets, we decide on the following hyperparameters: We employ 100-dimensional BPE and character embeddings, and the encoder and decoder hidden states are 300-dimensional. Dropout (Srivastava et al., 2014) is used with a probability of 0.5 for all hidden states when used as input to the next layer, as well as for the embedding layer. For training, we employ ADAM (Kingma and Ba, 2014). Whenever performance does not improve for 20 steps, we halve the learning rate and restart from the best performing model. Training stops when the learning rate gets below 0.0001; the best performing model is used for the final predictions. We do not use batching, since it hurts performance in our experiments on the development sets.

For decoding, we apply beam search with a beam of width 5.

4 Official System Evaluation

4.1 Datasets

The data for Task 2, track 2, LOW consists of sentences taken from the Universal Dependencies

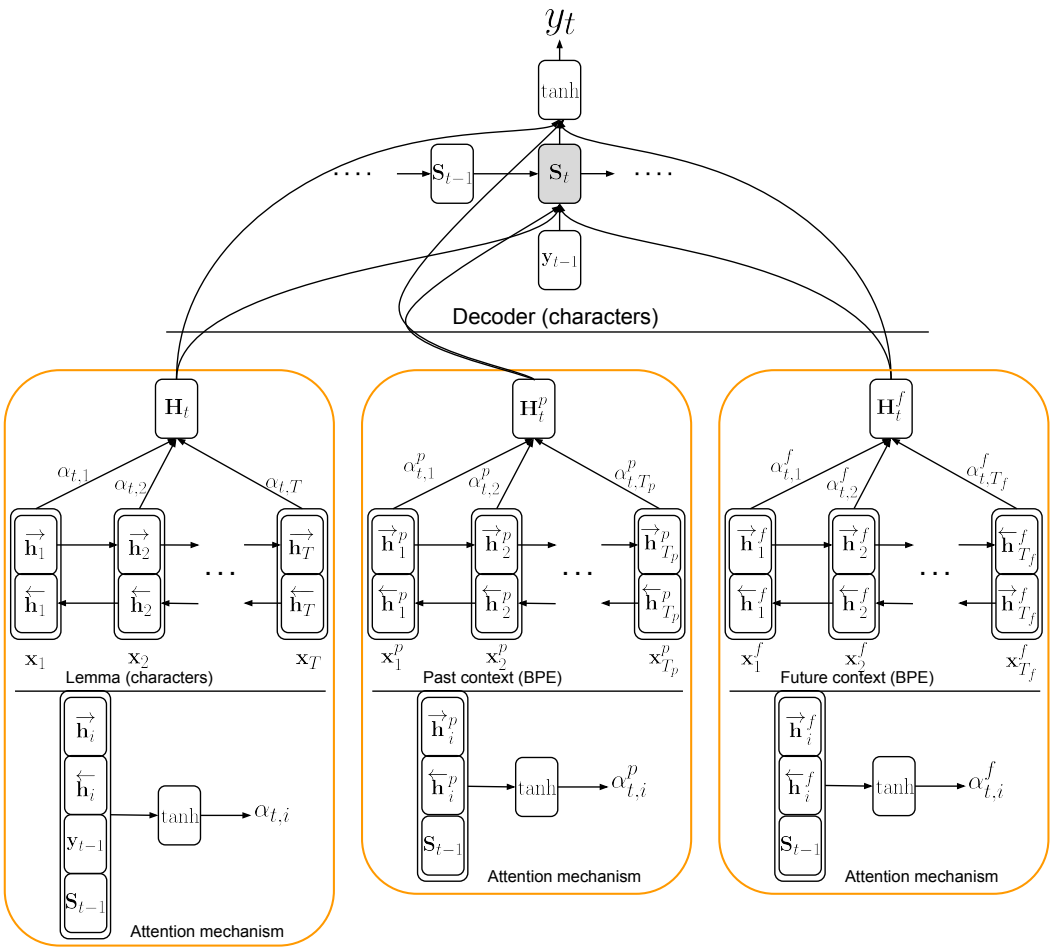

Figure 1: Overview of our employed model architecture.

(UD) treebanks (Nivre et al., 2017). All context forms, as well as the lemma of the target inflected form are given for each sentence. Training and development sets feature exactly one correct target form, while, for the test set, additional plausible target forms have been manually given by the shared task organizers (Cotterell et al., 2018).

The languages we experiment on are German, English, Spanish, Finnish, French, Russian and Swedish.

4.2 Baseline System

The official baseline system of the shared task is a character-level LSTM encoder-decoder model with attention (Bahdanau et al., 2015). The main input to the system is the lemma of the inflected form which is to be generated. Further, the context is taken into account: each character of the lemma is concatenated with 7 additional embeddings representing (i) the lemma of the word at the previous position in the sentence, (ii) the previous word itself, (iii) the tag of the previous word, (iv) the lemma of the word at the next position in the sentence, (v) the next word itself, (vi) the tag of the next word, (vii) the lemma of the inflected form to generate and given to the encoder. Note that, since no tags or lemmas are available for track 2 of Task 2, but the architecture is identical to that used for track 1 of the same task, all embeddings but those for the previous and the next word, as well as the lemma are set to default vectors.

Given the character embedding-context representations produced by the encoder, the LSTM decoder generates the character sequence of the output inflected form, using an attention mechanism.

More details on the shared task baseline system can be found in Cotterell et al. (2018).

4.3 Official Test Results

Two official results are reported. First, system performance is calculated by just taking the gold solution into account, i.e., all generated inflected forms that do not match the UD gold standard are counted as wrong. Second, performance is com-

	BL	BME-HAS	CPH	CUB	*NYU*	UZH
de	0.10	27.81	18.91	11.02	*44.08*	**59.15**
en	2.22	56.90	59.42	58.91	*66.57*	**68.08**
es	8.98	27.77	31.84	27.91	*27.49*	**32.68**
fi	0.38	8.89	12.33	7.88	*15.37*	**24.40**
fr	0.00	9.57	**29.53**	23.01	*26.48*	25.05
ru	2.71	19.68	22.69	21.08	*22.09*	**28.11**
sv	0.96	22.34	30.96	16.49	*31.60*	**32.77**
av.	2.19	24.71	29.38	23.76	*33.38*	**38.60**

Table 1: Test accuracies when considering only the gold solution; BL = BASELINE; CPH = COPENHAGEN; CUB = CUBoulder. Best results per language in bold; our results in italic.

	BL	BME-HAS	CPH	CUB	*NYU*	UZH
de	0.10	31.14	21.54	11.53	*48.43*	**61.38**
en	2.92	62.64	66.87	66.36	*72.21*	**74.02**
es	11.08	33.52	**37.31**	31.42	*31.98*	37.17
fi	0.89	11.18	16.14	10.04	*18.68*	**28.21**
ru	2.71	21.29	24.40	22.59	*23.29*	**30.42**
sv	0.96	27.34	36.38	19.04	*37.13*	**39.36**
av.	3.11	31.18	33.77	26.83	*38.62*	**45.09**

Table 2: Test accuracies when counting all plausible forms as correct; BL = BASELINE; CPH = COPENHAGEN; CUB = CUBoulder. Best results per language in bold; our results in italic.

puted by taking all plausible target inflected forms into account, i.e., all forms that *could be correct* in any way of reading the sentence are accepted as correct. The final results for all systems are shown in Tables 1 and 2, respectively.

As can be seen, the baseline performs poorly in the low-resource setting we consider here. In particular, its accuracy is far worse than that of any participating system.

Looking at our system's performance, we can see that it is the second best one for German, English, Finnish, French, and Slovene, as well as on average, when only considering the gold solution. Taking all plausible forms into account, our systems obtains the second highest accuracy for German, English, Finnish, and Slovene, as well as on average.[1]

The best performing system on average is UZH, and CPH outperforms our model for Spanish, French and Russian for gold solutions, and Spanish and Russian for all plausible forms. BME-HAS and CUB perform worse than our system for all languages.

A final observation is that the accuracy difference between the evaluation with the gold solution and the evaluation with all plausible forms is $0.92 - 6.49$, depending on the system.

[1]No results with all plausible forms are available for French.

5 Related Work

Most recent work on morphological reinflection was done in the context of the SIGMORPHON 2016 and the CoNLL–SIGMORPHON 2017 shared tasks.

The first edition of the shared task in 2016 (Cotterell et al., 2016) resulted in 3 different types of systems: "pipeline approaches" (unsupervised alignment algorithms applied to the source-target pairs, followed by a model which predicts edit operations), "neural approaches", and "linguistically inspired systems". The winning system was a neural network, namely a character-based RNN encoder-decoder model with attention, similar to the one we use here (Kann and Schütze, 2016). Hence, neural models gained popularity in the 2017 edition of the shared task (Cotterell et al., 2017a). In 2017, explicit low-resource settings were first introduced to the shared task. These settings demonstrated the effectiveness of hard attention in neural sequence-to-sequence models if training data are limited (Makarov et al., 2017).

Research not immediately done for the shared tasks included papers on multi-source reinflection (Cotterell et al., 2017b; Kann et al., 2017a), cross-lingual transfer for reinflection (Kann et al., 2017b), or first intents of neural inflection systems which make use of context for lemmatization (Bergmanis and Goldwater, 2018).

Older work on morphological inflection includes Ahlberg et al. (2014); Durrett and DeNero (2013); Nicolai et al. (2015); Faruqui et al. (2016), inter alia.

6 Conclusion

We presented the NYU system for Task 2, track 2, LOW of the CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. The system was designed for the task of morphological inflection in context: it predicts an inflected form for an indicated position in a sentence, given the sentence context and the lemma. In the official evaluation, which consisted of experiments in German, English, Spanish, Finnish, French, Russian and Slovene, our system was the second best performing one out of 5 submissions.

References

Malin Ahlberg, Markus Forsberg, and Mans Hulden. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *EACL*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with lematus. In *NAACL*.

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 *CoNLL–SIGMORPHON*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017a. The CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *CoNLL-SIGMORPHON*.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *SIGMORPHON*.

Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017b. Neural graphical models over strings for principal parts morphological paradigm completion. In *EACL*.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *NAACL-HLT*.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *NAACL-HLT*.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017a. Neural multi-source morphological reinflection. In *EACL*.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017b. One-shot neural cross-lingual transfer for paradigm completion. In *ACL*.

Katharina Kann and Hinrich Schütze. 2016. MED: The lmu system for the SIGMORPHON 2016 shared task on morphological reinflection. In *SIGMORPHON*.

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

Peter Makarov, Tatiana Ruzsics, and Simon Clematide. 2017. Align and copy: UZH at SIGMORPHON 2017 shared task for morphological reinflection. In *CoNLL-SIGMORPHON*.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *NAACL-HLT*.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebirolu Eryiit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çar Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökrmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà M, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phng Lê Hng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Lng Nguyn Th, Huyn Nguyn Th Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalnia, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.