Neural Morphological Segmentation for Polysynthetic Minimal-Resource Languages

Katharina Kann*
Center for Information and Language Processing
LMU Munich, Germany
kann@cis.lmu.de

Manuel Mager*
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
mmager@turing.iimas.unam.mx

Ivan Meza-Ruiz
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México
ivanvladimir@turing.iimas.unam.mx

Hinrich Schütze
Center for Information and Language Processing
LMU Munich, Germany
inquiries@cislmu.org

Abstract

In polysynthetic languages, a word can consist of many individual morphemes, which causes a strong need for morphological segmentation. However, many polysynthetic languages are *minimal-resource languages*, making the training of state-of-the-art neural segmentation systems difficult. Here, we present new morphological segmentation datasets for four indigenous Mexican languages, and show that neural sequence-to-sequence models obtain competitive performance even for small amounts of training data. Additionally, we introduce two novel data augmentation and two novel multi-task approaches, which further increase performance.

1 Introduction

Due to the advent of computing technologies to indigenous communities all over the world, natural language processing (NLP) applications for languages with limited computer-readable textual data are getting increasingly important. We aim at improving morphological surface segmentation—the task of splitting a word into the surface forms of its smallest meaning-bearing units, its *morphemes*—for these languages. Recovering morphemes provides information about unknown words, and is thus especially important for polysynthetic languages with a high morpheme-to-word ratio and a consequently large overall number of words.

Because of its relevance for linguistic analysis and down-stream tasks (Creutz et al., 2007; Dyer et al., 2008), segmentation has been tackled in many different ways (Creutz and Lagus, 2002; Ruokolainen et al., 2013, 2014). Recently, also neural approaches have been used, but mainly for canonical segmentation (Cotterell et al., 2016; Kann et al., 2016; Ruzsics and Samardzic, 2017). For surface segmentation, neural models have been used by Wang et al. (2016).

Here, we want to add two new questions to this line of research: (i) How can we successfully segment words in polysynthetic languages? (ii) Are neural networks applicable for morphological segmentation in minimal-resource settings and how can they be improved?

Our experiments show that neural sequence-to-sequence models perform roughly on par with strong state-of-the-art baselines for the polysynthetic languages Mexicanero, Nahuatl, Wixarika and Yorem Nokki in a minimal-resource setting. However, adding the multi-task and data augmentation methods which we will introduce in this work yields up to $5,05\%$ absolute accuracy improvement over our strongest baseline for 3 out of 4 languages.

*The first two authors contributed equally.

2 Polysynthetic Languages and Datasets

Polysynthetic languages are languages which are highly synthetic, i.e., single words can be composed of many individual morphemes. This property makes surface segmentation of polysynthetic languages especially complex but relevant for further linguistic analysis as well as down-stream tasks. We experiment on Mexicanero, Nahuatl, Yorem Nokki and Wixarika, which belong to the Yuto-Aztecan language family.

To create our datasets, we make use of both, words consisting of multiple morphemes and words consisting of one single morpheme, taken from books of the collection *archive of indigenous languages* (Canger, 2001; Lastra de Suárez, 1980; Gómez and López, 1999; Freeze, 1989). We first build test sets consisting of 40% of the available data, and then use 20% of the remaining instances to make the development sets. All other examples are used for training. We gathered a total of 1063 words for Yorem Nokki, 888 for Mexicanero, 1123 for Nahuatl, and 1394 for Wixarika.

3 Model and Extensions

Our approach is based on the neural sequence-to-sequence model introduced by Bahdanau et al. (2015) for machine translation. The first part of the model encodes the input sequence and consists of a bidirectional gated recurrent neural network (GRU) (Cho et al., 2014). Our input is the sequence of characters of the input word, represented by embedding vectors. The decoder is a single GRU, defining a probability distribution over strings in $\Sigma^* \cup \mathtt{S}$, for the language's alphabet Σ and a morpheme separation symbol $\mathtt{S}$. The probability of each new character is computed using an attention mechanism, and we employ an output softmax layer over $\Sigma \cup \mathtt{S}$.

Multi-task training. In order to leverage unlabeled data (MTT-U) or even random strings (MTT-R) during training, we employ multi-task training (Caruana, 1993) and define an autoencoding auxiliary task, which consists of producing an output which is equal to the original input string. We expect this to bias the model towards copying (the most frequent action it should take) and, in the case of MTT-U, to provide additional training data for the decoder's character language model.

Data augmentation. Another way to improve performance is to extend the available training data using unlabeled data or random strings. We build new training examples in a similar fashion as for the multi-task setup. All instances are of the form $w \mapsto w$, where either (i) $w \in V$ with V being words from a given unlabeled corpus (DA-U), or (ii) $w \in R$ with R being a set of sequences of random characters from the alphabet Σ of the language, i.e., $R \subset \Sigma^*$ (DA-R).

For both proposed methods, we treat the amount of additional training examples as a hyperparameter to be optimized on the development set.

4 Experiments and Results

Data. In addition to our datasets, for the multi-task training and data augmentation we use unlabeled data, collected from Gutierrez-Vasques et al. (2016), Mager Hois et al. (2017) and Maldonado Martnez et al. (2010).

Baselines. We compare our novel approaches to a fully supervised attention-based encoder-decoder RNN (Bahdanau et al., 2015) (S2S), the semi-supervised version of MORFESSOR (Kohonen et al., 2010) (MORF), and a strong discriminative conditional random fields model for segmentation by Ruokolainen et al. (2014) (CRFS).

Results. The final results of our proposed approaches as well as the baselines are shown in Table 1. It can be seen that S2S performs on par with CRFS for all languages but Nahuatl. S2S and CRFS both clearly outperform MORFESSOR (MORF).

All our proposed methods except for DA-U improve over S2S for all languages. DA-U, in turn, performs worse than S2S for all languages except for Mexicanero. This shows clearly that simple adding of corpus data confuses the model: it erroneously learns to not segment words that consist of multiple morphemes. Notably, this does not happen for random strings. We thus conclude that multi-task training (instead of simple data augmentation) is crucial for the use of unlabeled data, while random strings can be used for data augmentation as well as multi-task training.

Finally, with the exception of Nahuatl for which CRFS performs best, all of our novel methods achieve a higher accuracy than all baselines for all languages. This shows the effectiveness of our neural approaches for morphological segmentation in minimal-resource settings.

	MTT-U	MTT-R	DA-U	DA-R	S2S	MORF	CRFS
Mexicanero	**0.8051**	0.7955	0.7611	0.7983	0.7504	0.3364	0.7837
Nahuatl	0.6004	0.6027	0.5541	0.6018	0.5585	0.4044	**0.6444**
Wixarika	0.5895	0.6134	0.5425	**0.6188**	0.5754	0.3989	0.5866
Yorem Nokki	0.6856	**0.7101**	0.6212	0.6936	0.6569	0.4812	0.6596

Table 1: Accuracy of our approaches and baseline systems.

5 Conclusion

We investigated the applicability of neural encoder-decoder models to the task of surface segmentation for polysynthetic languages in minimal-resource settings. Our results showed that neural networks perform comparatively to or better than several strong baselines. We further proposed two novel multi-task approaches and two novel data augmentation methods to additionally increase performance on the task. For Mexicanero, Wixarika and Yorem Nokki our proposed methods outperform all baselines by up to 5.05% absolute accuracy.

References

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

Una Canger. 2001. *Mexicanero de la sierra madre occidental*. El Colegio de México.

Rich Caruana. 1993. Multitask Learning: A Knowledge-Based Source of Inductive Bias. In *ICML*.

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. In *EMNLP*.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. In *NAACL-HLT*.

Mathias Creutz, Teemu Hirsimäki, Mikko Kurimo, Antti Puurula, Janne Pylkkönen, Vesa Siivola, Matti Varjokallio, Ebru Arisoy, Murat Saraçlar, and Andreas Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *TSLP*, 5(1):3:1–3:29.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Workshop on Morphological and Phonological Learning*.

Christopher Dyer, Smaranda Muresan, and Philip Resnik. 2008. Generalizing word lattice translation. In *ACL-HLT*.

Ray A Freeze. 1989. *Mayo de Los Capomos, Sinaloa (Mayo of Los Capomos, Sinaloa).* ERIC.

Paula Gómez and Paula Gómez López. 1999. *huichol de san andrés cohamiata, jalisco*, volume 22. El Colegio de México.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *LREC*.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *EMNLP*.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. In *SIGMORPHON*.

Jesus Manuel Mager Hois, Meza Ruız, Ivan Vladimir, and Autónoma de México. 2017. Wixnlp: Probabilistic finite-state morphological analyzer for wixarika. In *WiNLP*.

Juan Pedro Maldonado Martnez, Crescencio Buitimea Valenzuela, and ngel Macochini Alonzo. 2010. *Vivimos en un pueblo yaqui*. SEP.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *CoNLL*.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *EACL*.

Tatyana Ruzsics and Tanja Samardzic. 2017. Neural sequence-to-sequence learning of internal word structure. In *CoNLL*.

Yolanda Lastra de Suárez. 1980. Náhuatl de acaxochitlán (hidalgo). *Archivos de lenguas indígenas de México. DF: Colegio de México*.

Linlin Wang, Zhu Cao, Yu Xia, and Gerard de Melo. 2016. Morphological segmentation with window lstm neural networks. In *AAAI*.