POSTERImproving Neural Morphological Segmentation for Polysynthetic Minimal-Resource Languages

Katharina Kann[1], Manuel Mager[2], Ivan Meza[2] and Hinrich Schütze[1]

[1] CIS, LMU Munich, Germany

[2]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas
Universidad Nacional Autónoma de México

Morphological Segmentation

The segmentation task aims to split a word into the surface forms of its smallest meaning-bearing units, its *morphemes*, i.e.:

ne|p+|ti|kuye|kai (*wixarika*)
I was sick (*English translation*).

Research Questions

1. How can we successfully segment words in polysynthetic languages?
2. Which supervised methods are applicable in minimal-resource settings and how can they be improved?

Polysynthetic Languages

Polysynthetic languages are languages which are highly synthetic, i.e., single words can be composed of many individual morphemes. We experiment on four languages of the Yuto-Aztecan:

- **Mexicanero**
- **Nahuatl**
- **Wixarika**
- **Yorem Nokki**

Improving Neural Seq2Seq

Multi-Task Training: We define an autoencoding auxiliary task, which consists of producing an output which is equal to the original input string, using Random String (MTT-R) and unlabeled words (MTT-U).

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{(w,c)\in\mathcal{T}} \log p_{\boldsymbol{\theta}}\left(c \mid e(w)\right) + \sum_{a\in\mathcal{A}} \log p_{\boldsymbol{\theta}}(a \mid e(a))$$

Data Augmentation: We extend the available training data with new examples from unlabeled data set (DA-U) and random strings (DA-R), such that $w \rightarrow w$.

Results

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

Model

Architecture: Attention-based encoder-decoder gated recurrent neural network (Bahdanau et al., 2015).

Hyperparameters: 100-dimensional hidden layers in encoder and decoder; 300-dimensional embeddings; training: stochastic gradient descent, Adadelta and minibatch size 20.

Data set

	train	dev	test	total
Yorem N.	511	127	425	1063
Mexic.	527	106	355	888
Nahuatl	540	134	449	1123
Wixarika	665	176	553	1394

Amount of Additional Data

We treat the amount of additional and artificial data as an hyperparameter. Values we experiment with are m times the amount of instances in the original training set, with $m \in \{1, 2, 4, 8\}$.

Conclusions

- We investigated the applicability of neural encoder-decoder models for surface segmentation.
- We proposed 2 novel multi-task approaches and 2 novel data augmentation methods.
- Our methods outperformed all baselines by up to 5.05% absolute accuracy in three languages.

Acknowledgements

To CONACYT (Program No. FC-2016-01-2225). We also thank the support of Gerardo Sierra.

Contact

Katharina Kann:
kann@cis.lmu.de
Manuel Mager:
mmager@turing.iimas.unam.mx