Reconciling Historical Data and Modern Computational Models in Corpus Creation

Joseph Rhyne
Cornell University

SCiL 2020

Introduction

Modern computational methods for corpus creation require more data and pre-tagged training sets. Historical data is extremely limited and usually not in model-ready form. How can we still accomplish the goal of corpus creation without time-consuming manual tagging? In this paper, we begin to answer this question.

Objectives

- Facilitate the creation of historical corpora to aid future analyses
- Extend modern models from related languages
- Approach historical data as low-resource language [1, 2, 3]

- Thousands of low-resource languages share this challenge, with computational approaches developed for them, e.g. use of parallel corpora. [2]
- **Model Transfer**[1]: Create cross-lingual embeddings [4] with **(1)** Bilingual dictionary and **(2)** two monolingual corpora to use with **(3)** small annotated corpus to tag texts.
- **Extending Modern Word Embeddings**: train models on the modern languages and extend them to the older stages and tag the historical texts.

Data: Old Slavic

Language	Pre-tagged	Untagged	Total
Old Church Slavonic	10	36	46
Old East Slavic	32	3	35
Old Polish	0	20	20

Model Transfer and Cross-Lingual Word Embeddings

From monolingual corpora, train monolingual embeddings. Use a bilingual dictionary to project both onto a common space as a cross-lingual embedding. [4] With a small tagged set, we can "transfer" the English model to the OCS texts.

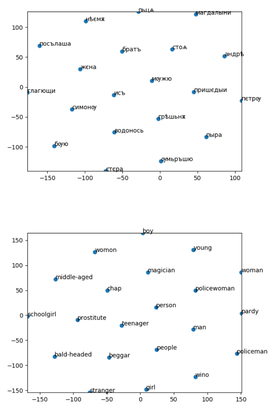


Figure: OCS and English embeddings near "man", plotted using t-SNE

```
жена - woman
водонось - vessel
грѣшьнѣ - sinner
исъ - Jesus
пьца - writing
умьрътвию - death
пришедыи - approach
боую - fool
пыра - dispute
єтера - certain
стоа - hundred
посълаша - messenger
магдалыни - Magdalene
глагающи - word
мужю - man
нъємѫ - mute
петрову - Peter
андрѣ - Andrew
симоню - Simon
братъ - brother
```

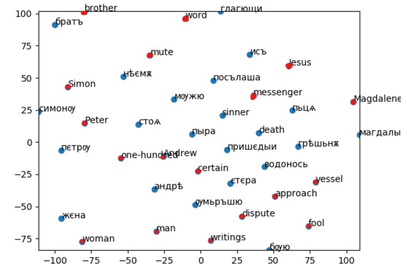


Figure: Cross-lingual embedding for OCS and English

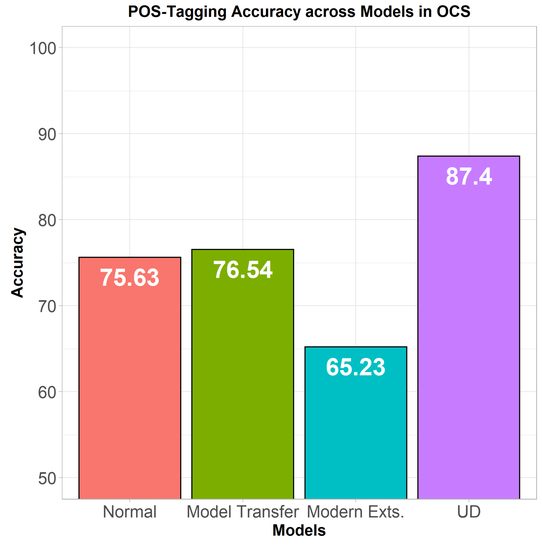
Results

All models use an extension of a sequence tagging network. [5] Embeddings trained using *word2vec*. [6]

- 3 Kinds of models, same test set for each:
  1. **Normal**: OCS and OES models trained on pre-tagged data
  2. **Model Transfer**: OCS, with OCS-English cross-lingual embedding
  3. **Modern Model Extensions**: UD models for Bulgarian, Russian, and Polish applied to the older related stages
- **UD models** applied to modern texts form a baseline comparison

Lang.	Normal	Mod.Exts.	UD
OES	69.60	70.95	83.91
OP	N/A	69.82	84.64

Figure: Accuracies for test set tagging in OES and OP



Conclusion

Some marginal improvements showing these models could be a good first-pass run for these and similar languages, considering their morphological complexity. Still many other low-resource methods to try.

Takeaways

1. **Model Transfer** and **Modern Model Extensions** can form the foundation of POS-tagging historical texts in corpus creation, augmented by manual annotation.

Selected References

[1] Meng Fang and Trevor Cohn.
Model transfer for tagging low-resource languages using a bilingual dictionary.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 587–593, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] Jan Buys and Jan A. Botha.
Cross-lingual morphological tagging for low-resource languages.
CoRR, abs/1606.04279, 2016.

[3] Pruthwik Mishra, Vandan Mujadia, and Dipti Sharma.
Pos tagging for resource poor indian languages through feature projection.
02 2018.

[4] Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith.
Massively multilingual word embeddings.
CoRR, abs/1602.01925, 2016.

[5] Nils Reimers and Iryna Gurevych.
Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 338–348, Copenhagen, Denmark, 09 2017.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean.
Distributed representations of words and phrases and their compositionality.
CoRR, abs/1310.4546, 2013.

Contact Information

- Web: cornell.academia.edu/JosephRhyne
- Email: jtr92@cornell.edu

Cornell University