Reconciling Historical Data and Modern Computational Models in Corpus Creation

Joseph Rhyne
Department of Linguistics
Cornell University
jtr92@cornell.edu

1 Overview

We live in a time of unprecedented access to linguistic data, from audio recordings to corpora of billions of words. Linguists have used these resources to advance their research and understanding of language. Historical linguistics, despite being the oldest linguistic subfield, has lagged behind in this regard. However, this is due to several unique challenges that face the subfield. Historical data is plagued by two problems: a lack of overall data due to the ravages of time and a lack of model-ready data that have gone through standard NLP processing. Barring the discovery of more texts, the former issue cannot be solved; the latter can, though it is time-consuming and resource-intensive. These problems have only begun to be addressed for well-documented language families like Indo-European, but even within these progress is slow.

There have been numerous advances in synchronic models for basic NLP tasks like POS and morphological tagging. However, modern models are not designed to work with historical data: they depend on large volumes of data and pre-tagged training sets that are not available for the majority of historical languages. Some have found success with methods that are designed to imitate traditional historical approaches, e.g. (Bouchard-Côté et al., 2013; McMahon and McMahon, 2003; Nakleh et al., 2005), but, if we intend to use state-of-the-art computational tools, they are essentially incompatible. This is an important challenge that computational historical linguists must address if they are going to meet the standards set by both modern corpora and historical analyses. This paper approaches the issue by treating historical data in the same way as a low-resource language (Fang and Cohn, 2017; Buys and Botha, 2016; Mishra et al., 2018) and integrating data from modern descendant languages. Through these approaches, we are able to tag a number of new texts in Old Slavic languages for part-of-speech. Many of these texts have never previously been tagged. With these problems overcome, we can create new corpora of historical language and thus dramatically increase both the number and type of diachronic linguistic investigations.

2 Modern approaches to historical data

Historical Data as low-resource language. This challenge is not unique to historical data. Thousands of languages across the world also lack the necessary resources for standard computational analyses and models. These low-resource languages have not been sufficiently documented and thus do not have adequate datasets for model-training. Many different approaches have been proposed on how to deal with this issue for low-resource languages. For example, (Buys and Botha, 2016) improve results through the use of parallel forpora, which could be helpful for those languages that have modern high-resource language translations. Others have proposed feature projection (Mishra et al., 2018) for morphologically-complex languages. In this paper, we exploit the approach called *Model Transfer* (Fang and Cohn, 2017). Here, a bilingual dictionary, monolingual corpora in both the high- and low-resource languages, and a small annotated corpus for the low-resource language, are used to train a model through joint training from both sources. The bilingual dictionary and monolingual corpora are used to train cross-lingual word embeddings, while language-dependent information can be learned from the small annotated corpus. The lack of available dictionaries for some languages is a pitfall for Model Transfer.

Extending modern language data. Historical data does not exist within a vacuum. One avenue that we could exploit is its relationship to descendant and related languages, i.e. how Modern English is a descendant of Middle English. We might leverage the large amount of pre-processed data available for the modern languages to help create the models for their older stages. We call this *Model Extension*, where a model is created to tag one language using training data from a related language. In this paper, we train models on modern data and use them to tag the older texts. Thus the model is extending to a new linguistic domain. No matter the approach, manual annotation is an option, and it goes a long way in helping to train models on these limited data.

3 Data

For this paper, we experiment on Old Slavic languages, focusing on Old Church Slavonic (OCS; 46 texts: 10 tagged, 36 untagged), Old East Slavic (OES; 35 texts: 32 tagged, 3 untagged), and Old Polish (OP; 20 untagged texts). These are good candidates because there are (1) resources for some of the languages (OCS and OES) and (2) well-documented modern descendant languages, i.e. Bulgarian for OCS, Russian for OES, and Polish for OP. Some pre-tagged texts for OCS and OES were taken from the TOROT treebank (Eckhoff and Berdiceviskis) to be used as training and test data. Untagged texts in all three languages were taken from sites like Thesaurus Indogermanischer Text- und Sprachmaterialien. OCS was the only language for which an extensive dictionary could be found, thus it is the only language to use *Model Transfer*. Word-embeddings were trained for the languages using the gathered texts. Models for the modern language were trained using data taken from Universal Dependencies.

4 Models

In order to tag the corpus we used an extension of a sequence tagging network, based on (Reimers and Gurevych, 2017) and (Arakelyan et al., 2018). These are based on BiLSTM networks from (Huang et al., 2015). For the models, we use a variety of both pre-trained embeddings for modern languages and newly-trained embeddings for the old languages, using Word2Vec (Mikolov et al., 2013). This set-up of the network can be seen in Figure 1.

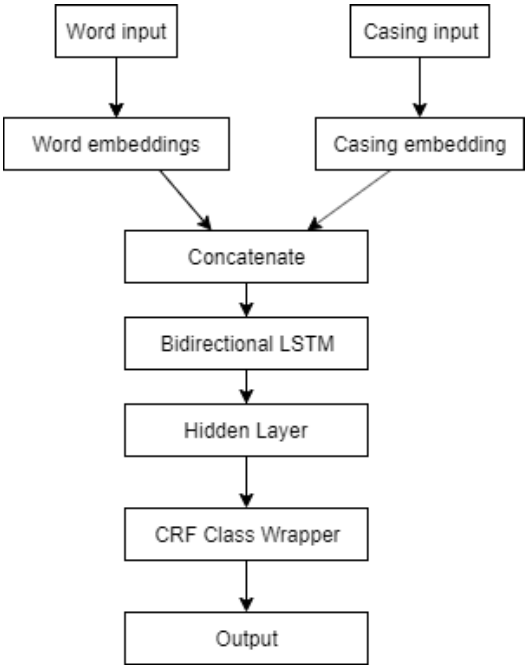

Figure 1: Basic architecture, showing the layers of the network used to create the models

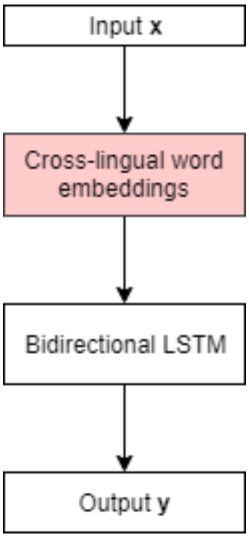

Figure 2: OCS Model Transfer has only one difference: the use of cross-lingual word embeddings (Ammar et al., 2016)

Based on this architecture, we trained three types of models: (1) Normal Models using the pre-tagged data for OCS and OES, (2) Model Transfer for OCS using an OCS-English dictionary and the British National Corpus, and (3) Modern Model Extensions using Universal Dependency models for Buglarian, Russian, and Polish. The OCS Model Transfer model had an additional requirement: following (Fang and Cohn, 2017), after the input of raw text, we use cross-lingual word embeddings (Ammar et al., 2016) instead of the usual

Language	*Normal*	*Model Transfer*	*Modern Extensions*	*Universal Dependency*
OCS	75.63	76.54	65.23	87.40
OES	69.60	N/A	70.95	83.91
Old Polish	N/A	N/A	69.82	84.64

Table 1: Accuracies for test set tagging in each language across different models

monolingual word embeddings. These combine monolingual embeddings trained using *word2vec* by projecting them onto a common space, which is learned through the bilingual dictionary. This is then used with the large tagged corpus of the high-resource language in the training, to then be applied to the untagged historical data. The tagging model itself uses the same BiLSTM architecture described above. The Model Transfer workflow can be seen in Figure 2. For comparison, Universal Dependency models were trained using the UD data for the three modern related languages: Bulgarian, Russian, and Polish.

5 Results

All models were subject to the same test set in each of the languages. Because there was no previously tagged corpus, the test set for Old Polish was hand-tagged for this project. This determined their POS tagging accuracies, which are compiled in Table 1.

None of the models achieved the same level of accuracy seen by the modern Universal Dependency models. The normally-trained models for OCS and OES were close, as a result of their pre-tagged data. In general, we can see that the use of Model Transfer and Model Extension does not negatively impact the POS tagging accuracy. The Extension model for OCS is lower than for the others, but this is likely due to dramatic morphological differences between OCS and its modern relative Bulgarian. While the overall accuracies are not as high as most modern language models, they are not so low as to be discouraging. They do show that, in the instance of a language like Old Polish, Model Transfer and Extension are serviceable methods for tagging new texts. Even at a 70% POS-tagging accuracy, these methods provide a great first-pass run in the pipeline of corpus creation for a language without resources. Moreover, this maintenance of a comparably high accuracy shows that we can leverage different stages of a language to fill in gaps in our models. This is still likely dependent on other diachronic factors, e.g. we might expect a lower accuracy for an older morphologically-complex language when its descendant form is much more morphologically-simple.

6 Conclusion

The results so far do not meet the standards set by modern models, but they do still serve as a good first-pass run that can be improved with manual annotation and other tagging methods. This will still save valuable time and increase the number and type of resources available to historical linguistics. This in turn will further aid historical linguists in both their diachronic and synchronic analyses for the languages and language families included in the new corpora, e.g. (Rhyne, Forthcoming). This can only improve with access to more data. Nevertheless, it is still promising that models can be extended relatively well from modern languages to their ancestors. Moreover, there are still multiple low-resource language approaches that can still be used, such as parallel corpora (Buys and Botha, 2016). This would be especially useful for languages that have extensive English or other modern translations. We might also try to use dictionaries of modern descendant languages in our Model Transfer approach.

Thus, this paper attempts to fill in a gap that continues to plague historical linguistics. The results are still lacking, but they show signs of improvement. With more time and resources, other methods could be explored, particularly those that depend on extensive pre-tagged data. Nevertheless, through efforts like these, we can improve the quality of data within historical linguistics, making it more approachable to all linguists and matching the standards already established in the rest of the field.

References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith.

2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Gor Arakelyan, Karen Hambardzumyan, and Hrant Khachatrian. 2018. Towards JointUD: Part-of-speech tagging and lemmatization using recurrent neural networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 180–186, Brussels, Belgium. Association for Computational Linguistics.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, , and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110:4224–4229.

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

Hanne Martin Eckhoff and Aleksandrs Berdiceviskis. Linguistics vs. digital editions: The Tromsø Old Russian and OCS Treebank. In *Scripta and e-Scripta 14-15*, pages 9–25.

Meng Fang and Trevor Cohn. 2017. 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. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

April McMahon and Robert McMahon. 2003. Finding families: quantitative methods in language classification. *Transactions of the Philological Society*, 101(1):7–55.

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

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

Luay Nakleh, Tandy Warnow, Don Ringe, and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an indo-european dataset. *Transactions of the Philological Society*, 103:171–192.

Nils Reimers and Iryna Gurevych. 2017. 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.

Joseph Rhyne. Forthcoming. Contrasts in case usage under negation in old church slavonic. In *Proceedings of the 29th Annual UCLA Indo-European Conference*, Bremen. Hempen.