Expanding morphosyntactic resources for Ukrainian from cognate Slavonic languages

Abstract

This paper describes a method for developing resources for POS tagging and dependency parsing for Ukrainian from resources available for other Slavonic languages. This is the first publicly available model that provides accurate POS tagging.

1 Introduction

Tools for POS tagging and parsing are among the most basic tools needed for Natural Language Processing. The modern method of their development implies training generic tools on a large sample of annotated materials.

The annotation scheme can vary from the Penn tagset, commonly used for processing English, to Multext, which covers a wide range languages (Erjavec, 2010) to the Universal Dependencies (UD) tagset (Nivre et al., 2016).

The tagsets can be usually mutually converted, even with some loss of information if a more specific category is mapped to a less specific one (Zeman, 2008).

Variations in tagsets: participles as verbs or as adjectives. In conversion from the MTE tagset into the UD one, information about the case governing model for the prepositions is lost, `Sp-g` in Table 1 indicates that the preposition in this sentence expects the genitive case of the following noun phrase.

While some languages within the existing training resources in the UD set have sizeable UD treebanks for training, e.g., nearly 2 million tokens for Czech and 1 million for Russian, there are little or no training resources available for other languages, including Ukrainian, see Table 2.

Empirical investigation on the advantages of noisy sources (Hovy et al., 2014).

2 Methodology

- translation of cognate words into Ukrainian (Babych, 2016) (**by November 2016**);
- conversion of existing UD corpora into Ukrainian (Reddy and Sharoff, 2011) (**by December 2016**);
- correcting the biases for systemic morphological mismatches (**by February 2017**);

For training we experimented with several frameworks:

- TnT, a classic HMM tagger (Brants, 2000);
- RF Tagger, a decision-tree approach to large tagsets (Schmid and Laws, 2008);
- Marmot, a CRF-based tagger (Müller et al., 2013);
- UDpipe, a perceptron-based tagger and transition parser (Straka et al., 2016);
- MaltParser, an SVM-based transition parser (Nivre et al., 2006)

In addition we wanted to investigate the influence of the typological language distance. Typologically Ukrainian is in the group of the East Slavonic languages. However, very substantial resources available for Czech (a West Slavonic language) make it interesting to investigate the effect of having a larger more distant set in comparison to a smaller closer one.

Even if the tagging resources are not available, some lexicons with morphological annotation are available (Derzhanski and Kotsyba, 2009). They can be integrated into the frameworks listed above to reduce the Out-Of-Vocabulary (OOV) rate.

Systemic morphological mismatches can be corrected by comparing the frequency distribution of tagged and parsed wikipedias in the donor and recipient languages.

Word	MTE	UD tag
делается	Vmip3s-m-e	VERB|Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
из	Sp-g	ADP
стали	Ncfsgn	NOUN|Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing

Table 1: MTE and UD tagsets compared

Table 2: Size of corpora in tokens

Languages	Training	Wikipedia
Bulgarian	160K	55M
Croatian	82K	40M
Czech	1954K	110M
Polish	90K	227M
Russian	946K	420M
Slovenian	158K	321M
Ukrainian	-	161M

3 Experiment

3.1 Sources of data

In this study we used UD datasets for the Slavonic language family, as well as the respective Wikipedias, in order to build resources for processing Ukrainian.

We might also try Czech→Polish.

4 Results

By March 2017

Baseline accuracy for tagging and parsing: Czech and Russian, this is a bit lower than what is reported in other studies, namely (Sharoff and Nivre, 2011; Straka et al., 2016), possibly because of the tagset conversion (??for Czech).

Comparing various transfer options: $ru \rightarrow uk$ vs $cd \rightarrow uk$ vs $cs + ru \rightarrow uk$ vs $cs + ru + uk_{morph} \rightarrow uk$

Time for some linguistics?? (Manning, 2011), error analysis

5 Conclusions and future work

Not limited to Ukrainian.

A link between rules and Machine Learning

References

Babych, B. (2016). Graphonological levenshtein edit distance: Application for automated cognate identification. In *Proc EAMT*.

Brants, T. (2000). TnT - a statistical part-of-speech tagger. In *Proc. of 6th Applied Natural Language Processing Conference*, pages 224–231, Seattle.

Derzhanski, I. and Kotsyba, N. (2009). Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian. In *Proc Mondilex Third Open Workshop*, pages 9–26.

Erjavec, T. (2010). Multext-East Version 4: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc LREC*, Valletta, Malta.

Hovy, D., Plank, B., and Søgaard, A. (2014). When pos data sets don't add up: Combatting sample bias. In *Proc LREC*, pages 4472–4475.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational Linguistics and Intelligent Text Processing*, pages 171–189. Springer.

Müller, T., Schmid, H., and Schütze, H. (2013). Efficient higher-order crfs for morphological tagging. In *Proc EMNLP*, pages 322–332.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proc LREC 2016*.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proc LREC*, pages 2216–2219, Genoa.

Reddy, S. and Sharoff, S. (2011). Cross language POS taggers (and other tools) for Indian languages: An experiment with Kannada using Telugu resources. In *Proc IJCNLP'11*, Chiang Mai, Thailand.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proc COLING*, Dublin.

Sharoff, S. and Nivre, J. (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proc Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo.

Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proc LREC 2016*.

Zeman, D. (2008). Reusable tagset conversion using tagset drivers. In *Proc LREC*.