Building a Dependency Parsing Model for Russian with MaltParser and MyStem Tagset

Kira Droganova

Faculty of Humanities
Higher School of Economics
E-mail: kira.droganova@gmail.com

Abstract

The paper describes a series of experiments on building a dependency parsing model using MaltParser, the SynTagRus treebank of Russian, and the morphological tagger Mystem. The experiments have two purposes. The first one is to train a model with a reasonable balance of quality and parsing time. The second one is to produce user-friendly software which would be practical for obtaining quick results without any technical knowledge (programming languages, linguistic tools, etc.).

1 Introduction

There was a number of experiments on building dependency parsing models for Russian using MaltParser conducted previously. MaltParser was suggested and described by Nivre (Nivre et al. [1]). They did not include Russian in the languages used for experiments when describing general methodology and evaluation, however the subsequent experiments were performed on the SynTagRus treebank of Russian (Nivre et al. [2]) which currently contains 860 000 words. During the training of the model both lexical and morphological features were used. Further work presented by Sharoff (Sharoff, Nivre [3]) describes pipeline and tools for processing Russian texts. This software is represented as a set of scripts which need to be put together before use. All previously reported experiments were carried out involving TnT for POS-tagging and MaltParser for syntactic parsing.

In our approach, we use morphological information as the only input. The morphological tagger Mystem (Segalovich [4]) was designed specifically for Russian and has extremely useful settings which allow to make disambiguation by context. Moreover, the original morphological tagset of SynTagRus, ETAP-3 (Iomdin et al. [5]), is closer to Mystem than to TnT tagger. Since this has a direct influence on the quality of the parsing, our experiments was conducted using Mystem.

2 Approach

The project was divided into three levels: POS-tagging, training data, and tuning MaltParser settings. The final models were trained by combining the best results for each level. Thus, the pipeline was:

1. To prepare Mystem annotation using SynTagRus.

Mystem annotation were obtained using two methods. The original tagset of SynTagRus was mapped to the Mystem tagset using a conversion table in order to improve the accuracy of the tagging. There are certain mismatches in Mystem and ETAP3 tagsets, for example, personal pronouns are tagged as nouns in SynTagRus, there is no predicatives, parentheses and some other POS-tags. Moreover, SynTagRus includes multitokens (multi-word prepositions, adverbs, etc.). All these variations could affect parsing quality relating to actual data.

An alternative approach is to re-annotate SynTagRus, i.e. to get a certain word form from SynTagRus and send it to Mystem. The result is generally more accurate, but the composites (e.g. general-major 'Major General') often get recognized as two separate tokens by Mystem and as one token by ETAP 3, and vice versa, and this results in erroneous output. Therefore at present the quality of the models is much worse compared to the first approach.

In the future we are planning to apply Mystem (a version with disambiguation) directly both during training and during annotation of new data. We expect the reannotation to help to produce more accurate tags for composites during training and obtain better results.

2. To prepare training data by converting SynTagRus into conll-file.

SynTagRus was split into three parts: the training set (80%), the development test set (10%) and the final test set (10%). The original SynTagRus format (Iomdin et al. [5]) was converted into conll-file [6] using a convertion scheme.

Figure 1 provides an example of conversion scheme. Lines 1 and 3 provide information about SynTagRus structure, lines 2 and 4 relate to conll layer, which is the data format for MaltParser. The conversion scheme was developed for the purpose of transforming SynTagRus data into training data in conll format. For example, value " root" of the attribute "DOM" indicates the head of the sentence and should be converted into zero in the 6th conll layer position and into "root" in the 7th position. First three positions are typically converted inalterably. Concerning conll layer positions from 4 to 6, variations are allowed, such as for instance, part of speech and morphological data separation.

There was a number of experiments on a size of the data, punctuation marks and content of a field [6] performed previously. The most valuable experiments were performed on CPOSTAG (Coarse-grained part-of-speech tag) and POSTAG (Fine-grained part-of-speech tag) fields, where the 'three letter models' were trained. These models have three letters from the word ending in CPOSTAG or POSTAG.

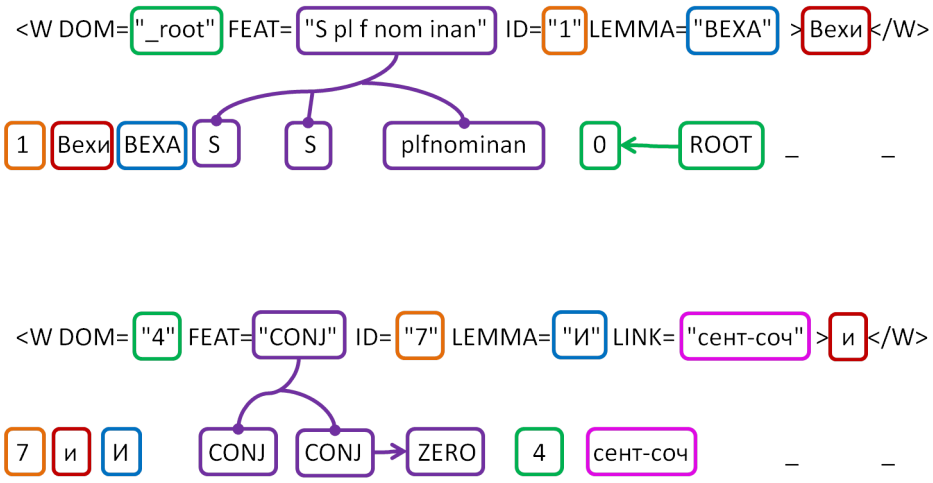

Figure 1: An example of conversion scheme

3. To train the models with different settings.

A large number of experiments was conducted using different types of projective and non-projective algorithms. Non-projective algorithms allow branches to cross as opposed to projective ones. The most valuable results have been achieved while using pseudo-projective transformations provided by MaltParser functionality.

3 Results

Results were measured with MaltEval using Labeled Attachment Score and Unlabeled Attachment Score evaluation metrics [7]. Accuracy reaches 80.3% by LAS and 87.5% by UAS for the best model with punctuation. Error evaluation is based on approach described by Toldova [8] adapted for the purpose of these experiments. The classification comprises 5 error types:

- Type 1 — wrong root predicted.
- Type 2 — wrong head predicted.
- Type 3 — wrong label predicted.
- Type 4 — wrong head predicted (acceptable error).
- Type 5 — wrong label predicted (acceptable error).

Type 1 is common for compound sentences longer than ten words. Normally, if the sentence has type 1 error, it has many type 2 errors. A large amount of

type 3 errors is due to special aspects of syntactic relations in Russian. 65 types of syntactic relations are used in SyntagRus and this results in lack of examples for some rarely used relations. Due to the so-called 'free word order' in Russian, wrong labels appear almost every time the head is predicted incorrectly. There are however more reasons for false labeling.

Types 4 and 5 are not indicative as they do not have significant effect on parsing result. These errors appears when individual type of syntactic relation is predicted as a general type or when predicted relation is an alternative version of tagging.

Future work includes a deeper analysis of the training data (word frequency, uncommon words), experiments with transforming syntactic relations into more simple structure and with special attention to the universal dependencies.

4 Conclusion

The paper presents the first results of building the dependency parsing model as the first step to produce a digital resource for linguists. Using all of the original SynTagRus syntactic relations and Mystem POS-tagging the model accuracy reaches up to 80.3% by LAS and 87.5% by UAS. Even though the results reported by Sharoff and Nivre [3] are slightly better (for SynTagRus tags: LAS 83.4, UAS 89.4), they are not comparable to ours due to the differences in training data and impossibility to replicate the experiments on the same dataset.

5 Acknowledgements

The author is grateful for computational capabilities provided by Andrey Kutuzov and mail.ru group.

References

[1] Nivre, Joachim, Hall, Jonah, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandra, Marinov, Svetoslav, Marsi, Erwin (2007) MaltParser: A language independent system for data-driven dependency parsing. Natural Language Engineering, 13, pp. 95-135.

[2] Nivre, Joachim, Boguslavsky, Igor M., Iomdin, Leonid L (2008) Parsing the SYNTAGRUS Treebank of Russian, In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pp. 641–648.

[3] Sharoff, Serge, Nivre, Joachim (2011) The proper place of men and machines in language technology. Processing russian without any linguistic knowledge', In Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference Workshop Dialogue 2011). Vol. 10 (17), 2011. Moscow: RGGU, pp. 657-670.

[4] Segalovich, Ilya (2003) A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In MLMTA-2003. (URL: https://tech.yandex.ru/mystem/)

[5] Iomdin, Leonid, Petrochenkov, Vyacheslav, Sizov, Viktor, Tsinman Leonid (2012) ETAP parser: state of the art. In Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2012. Vol. 11 (18), 2012. Moscow: RGGU, pp. 830-848.

[6] Depparse Wiki (URL:http://depparse.uvt.nl/DataFormat.html)

[7] Nilsson, Jens, User Guide for MaltEval 1.0 (beta), 2014 (URL:http://www.maltparser.org/malteval.html)

[8] Toldova, Svetlana, Sokolova, Elena, Astafiyeva, Irina, Gareyshina, Anastasia, Koroleva, Anna, Privoznov, Dmitry, Sidorova, Evgenia, Tupikina, Ludmila, Lyashevskaya, Olga (2012) Ocenka metodov avtomaticheskogo analuza teksta 2011-2012: Sintaksicheskie parsery russkogo jazyka [NLP evaluation 2011-2012: Russian syntactic parsers]. In Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue 2012. Vol. 11 (18), 2012. Moscow: RGGU, pp. 797-809.