Converting Russian Dependency Treebank to Stanford Typed Dependencies Representation

Janna Lipenkova
Institut für Deutsche und
Niederländische Philologie
Freie Universität Berlin
janna.lipenkova@fu-berlin.de

Milan Souček
Lionbridge Technologies, Inc.
Tampere, Finland
milan.soucek@lionbridge.com

Abstract

In this paper, we describe the process of rule-based conversion of Russian dependency treebank into the Stanford dependency (SD) schema. The motivation behind this project is the expansion of the number of languages that have treebank resources available in one consistent annotation schema. Conversion includes creation of Russian-specific SD guidelines, defining conversion rules from the original treebank schema into the SD model and evaluation of the conversion results. The converted treebank becomes part of a multilingual resource for NLP purposes.

1 Introduction

Dependency parsing has provided new methods and resources for natural language technology tasks in recent years. Dependency treebanks are now available for many languages and parsed data is used for improving machine translation, search engines and other NLP applications. While data resources are relatively common for monolingual tasks, there is a growing need for consistently annotated multilingual data. First larger activities in generating comparable standardized sets of multilingual parsed data were presented in CONLL shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007; Surdeanu et al., 2008; Hajič et al., 2009). More recently, cross-language consistency has been achieved by using one universal schema for all covered languages (McDonald et al., 2013). This universal treebank schema uses consistent sets of part-of-speech (POS) (Petrov et al., 2012) and dependency labels (deprel) following the Stanford typed dependencies representation (SD) (de Marneffe and Manning, 2008). The consistent treebank schema has many advantages, mainly the more straightforward possibility to build applications for multiple languages (McDonald et al., 2011), though it also presents challenges such as handling language-specific features among languages of different types without introducing conflicts. A certain level of generalization of language features that might lead to simplification is needed, as already highlighted by McDonald et al. (2013). For such universal multilingual treebank model, more resources can be built manually or they can be obtained by converting existing treebanks that follow different parsing schemas into one consistent treebank model. For the SD schema, treebanks for several languages already have been built using the manual annotation procedures (McDonald et al, 2013; Souček et al., 2013). There are also other existing treebanks covering languages from different families where the SD schema was applied (e.g. Chang et al., 2009 for Chinese; Haverinen et al., 2010 for Finnish; Seraji et al., 2012 for Persian; Tsarfaty, 2013 for Hebrew). Treebank conversion was applied e.g. in Italian (Bosco et al, 2013). The conversion model is a more suitable option for languages for which treebanks are already available, since manual annotation can be limited and the annotation/conversion process can be automated. In this paper, we describe the conversion of an existing Russian treebank

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 143–147,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

(SynTagRus, Boguslavsky et al., 2000) into the SD schema. We present the conversion process (2), introduce the source and the target model, including adaptations of the SD schema for Russian (3), describe the conversion script (4) and finally compare the conversion results (in terms of process efficiency and output accuracy) to other tasks for which a similar process was applied (5).

2 Conversion process

The conversion procedure used in our project is similar to the process described in Bosco et al. (2013). The very first step was the development of the Russian-specific set of POS and the list of Stanford dependency relations, compliant with the standards presented by Petrov et al. (2012) and de Marneffe and Manning (2008). Next, we investigated POS and deprel sets for the original treebank and defined conversion rules for adapting data to our specific schema. A rule-based conversion script was developed on a small portion of the original data and the remaining data was automatically converted. The quality of the conversion output was monitored by manual review of samples of converted data and also using parser evaluation methods. During the manual review, errors in samples of converted data were manually corrected in order to produce a gold standard data set; at the same time further conversion rules were reported for improving the conversion script. This cycle was repeated several times until an acceptable output quality was reached.

3 Source and target models

3.1 Source model

The source data for Russian are taken from the SynTagRus treebank (Boguslavsky et al., 2000). Just as in the basic SD model, SynTagRus provides a dependency tree for each sentence where each word is connected to a head and assigned one of 78 deprels, theoretically motivated by Melcuk's Meaning-Text theory (Melcuk, 1981). Additionally, the treebank specifies POS information as well as applicable morphological information (gender, number, case, degree of comparison, aspect, tense, person, voice).

3.2 Target model

The basic version of SD (de Marneffe and Manning, 2008) counts approximately 53 dependency labels. They are used in conjunction with a “universal” set of part-of-speech labels (Petrov et al., 2012). Although our aim is to build a resource that follows a consistent schema with other existing SD languages, we decided to make some minor modifications to the SD model to account for language-specific phenomena and thus minimize the loss of structural information. Both the set of SD dependencies and of POS labels were slightly adjusted to adapt the model to Russian. All these specifics can be further converted to the fully consistent SD model. The following modifications were made to the dependencies annotation schema:

- *scomp* is introduced for the complements of (ellipted) copulas.

- *ocomp* is introduced for verb complements that are semantically predicated by the object of the verb (e.g., *I find [this idea]$_i$ interesting$_i$.*).

- *gmod* is introduced for genitive modifiers of nominals; in turn, the *poss* relation for prenominal possessive modifiers is eliminated.

- *interj* is introduced for discourse particles attaching to nominals or verbs.

Despite the modifications, the adopted model still leads to losses in more fine-grained information. An example where this becomes especially visible are objects of verbs: the SD model uses the two labels *dobj* and *iobj* for direct and indirect objects. In Russian, there is a larger range of object types; they are distinguished not only morphologically, but also syntactically (e.g. genitive of negation, whereby the negation of the verb leads to the `switch' of the direct object from accusative to genitive). In order to capture these distinctions, the original treebank uses five relations (*1-compl*, *2-compl* etc.). However, the reduction to the two types *dobj* and *iobj* assumed for our SD model `deletes' these more fine-grained distinctions.

4 Conversion Script

4.1 General approach

The conversion script works with conversion patterns, which specify the possible targets of a source label and the conditions for this target to be applied. Conditions can be specified in terms of POS, morphological, lexical and structural information. Most conversion patterns have a regular formal basis and can be grouped into data structures that are processed by standardized functions. However, there are many patterns, especially less frequent ones, that have an irregular, more complex structure and thus have to be described individually. In order to increase the flexibility in formulating conversion patterns by specifying lexical information, the script is enriched with a set of lexical lists. These lexical lists mostly contain closed classes of functional words or idiosyncratic items, such as pronouns, subordinating conjunctions or idioms.

4.2 Conversion

Conversion acts on three types of information – POS tags, dependency relations and tree structures.

4.2.1 POS tags

The original data are annotated with five POS labels (NOUN, VERB, ADJ, ADV, CONJ); the target set contains 15 POS labels. One-to-many correspondences, for example the ambiguity of original NOUN between target NOUN and PRON, mostly occur in cases where the original POS tag subsumes some smaller class of functional words. As described above, word lists were used to identify these closed classes and to choose between the possible target POS tags.

4.2.2 Dependency relations

In the original treebank, 78 dependency relations are used; the target model contains 51 relations. For some original dependency labels, a one-to-one correspondence can be established. For example, the original label *advrb-subj*, used for nominals with an adverbial function, is always converted to *npadvmod*. However, most original dependency labels have multiple SD counterparts; conditional branching is used to determine the target relation for a given case. All types of information available in the treebank – POS, morphological, lexical and structural information – can be used to formulate conditions; in most cases, the specification for a given source relation involves a mix of the different information types.

Examples for the different types of conversion conditions are as follows:

- POS tag condition: *attrib*: convert to *nn* if NOUN, *amod* when ADJ, *prep* when ADP
- Morphological condition: *aux*: convert to *npadvmod* if in ablative case, *iobj* if in dative case.
- Structural condition: *explet*: convert to *mark* if dependent of *purpcl* or *rcmod*; *ccomp* if dependent of *complm*.
- Lexical condition: *aux*: convert to *neg* if expressed by *не/ни*, else *interj.*

4.2.3 Structural modifications

Structural modifications were introduced in several cases; most of them are caused by the reliance of SD on semantic heads and, thus, on content words as heads. During conversion, head-dependent structures are “switched” in cases where the original head does not correspond to the “semantic” head. Specifically, this occurs in the following cases:

- Structures with auxiliary verbs (future tense, passive voice): switch between auxiliary and lexical verb, so that the auxiliary depends on the lexical verb.
- Clauses introduced by subordinating conjunctions: switch between introducing conjunction and verb in the subordinate clause, so that the verb in the subordinate clause depends

no more on the conjunction, but on the verb in the matrix clause.

- Coordination structures: in the original data, the conjuncts and coordination particles form a chain, whereby each node depends on the previous one. In the target representation, all conjuncts and coordination particles attach to the first conjunct.

4.2.4 Problems and inaccuracies - syntactic under-specification

Under the SD model, different dependency relations may apply to structurally identical, but semantically different relations. For example, postverbal nominals in instrumental case can be either *iobj* (indirect object, corresponding to the instrument argument) or *npadvmod* (nominal adverbial); the relation applicable in a given case depends on the lexical semantics of the verb and the nominal:

(1) a. *npadvmod*(gaze, wolf):
смотреть *волком*
gaze wolf.INS
'to gaze angrily'
b. *iobj*(cut, knife):
резать *ножом*
cut knife.INS
'to cut with a knife'

The semantic difference is not visible at a surface level: there is no structural criterion which might condition the choice of the target relation. Since both structures are lexically productive, basing the choice on word lists is also not a satisfactory solution. Rather, the disambiguation of these and similar cases would require a more fine-grained semantic classification specifying valence frames and selectional restrictions of verbs as well as semantic features of nouns; in example (1), such a classification would allow to identify verbs that semantically select instruments (corresponding to *iobj*) as well as nouns that can potentially act as instruments. Besides, machine learning techniques can also be used for disambiguation based on the frequency of the lexical constellations for a particular dependency relation. Another problem are non-frequent dependency relations and contexts of occurrence which do not provide enough evidence for postulating a reliable, universally applicable conversion pattern. In the original treebank, 24 out of 78 dependency relations have a frequency of occurrence of less than 100. Besides, after the application of the conversion patterns, numerous dependency relations remain non-converted, because their contexts of occurrence are non-frequent and thus also cannot be reliably captured by conversion patterns. Our model uses the generic label *xdep* to identify tokens for which conversion was not successful. This label mostly appears for tokens whose original deprels do not allow for a rule-based characterization because they are partially defined in semantic terms, such as *nonself-agent*, *distrib*, *elaborat* and *mod-descr*.

5 Results

The presented script converts the original Russian treebank fully into the SD schema. The converted treebank data is owned by Google and its availability can be checked with the data owners. Conversion output precision was measured with MaltEval (Nilsson and Nivre, 2008) using manually annotated 500 sentences as gold standard and the same set processed with the conversion script as a test data. We achieved 76.21% LAS and 83.84% UAS. Achieved LAS is slightly lower than for similar work reported for Italian (Bosco et al., 2013), where LAS for different sub-models is between 79.94% and 84.14% in the parser output. Since the aim of this project is to create comparable cross-language data with acceptable precision within reasonable time frame, the precision that we achieved seems to be in acceptable range for the described conversion task.

6 Conclusions and future work

Our further target is to build similar conversion tasks for other languages, where existing treebanks are available. We also plan to take advantage of machine learning mechanisms that can make the conversion work more efficient.

Acknowledgemnts

This work was performed by Lionbridge NLS team for Google. We would like to thank Google for giving us the opportunity to work on this project and to publish our results.

References

Igor Boguslavsky; Svetlana Grigorieva; Nikolai Grigoriev; Leonid Kreidlin; Nadezhda Frid. 2000. *Dependency Treebank for Russian: Concept, Tools, Types of Information.* In: COLING 2000 Volume 2.

Cristina Bosco, Simonetta Montemagni, Maria Simi. 2013. *Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank.* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).

Sabine Buchholz and Erwin Marsi. 2006. *CoNLL X shared task on multilingual dependency parsing*. In Proceedings of CoNLL.

Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. *Discriminative reordering with Chinese grammatical relations features.* In Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antónia Martí, Lluís Márquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, Yi Zhang. 2009. *The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages*. In Proceedings of CoNLL.

Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. *Treebanking Finnish*. In Proceedings of TLT9, pp. 79–90

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. *Stanford typed dependencies manual.*

Ryan McDonald, Slav Petrov, and Keith Hall. 2011. *Multi-source transfer of delexicalized dependency parsers*. In Proceedings of EMNLP.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. *Universal Dependency Annotation for Multilingual Parsing.* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).

Igor Melcuk. 1981. Meaning-Text Models: A recent Trend in Soviet Linguistics. Annual Review of Anthropology 10, 27–62.

Jens Nilsson, and Joakim Nivre. 2008. MaltEval: An Evaluation and Visualization Tool for Dependency Parsing. In Proceedings of LREC.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. *The CoNLL 2007 shared task on dependency parsing*. In Proceedings of EMNLPCoNLL.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In Proceedings of LREC.

Mojgan Seraji, Be´ata Megyesi, and Nivre Joakim. 2012. *Bootstrapping a Persian dependency treebank*. Linguistic Issues in Language Technology, 7.

Milan Souček, Timo Järvinen, Adam LaMontagne. 2013. *Managing a Multilingual Treebank Project.* In Proceedings of the Second International Conference on Dependency Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. *The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies*. In Proceedings of CoNLL.

Reut Tsarfaty. 2013. *A unified morpho-syntactic scheme of Stanford dependencies*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).