Non Standard Treebank Parsing

Victoria Bobicev, Cătălina Mărănduc

Abstract

We report on our ongoing work on Dependency Treebank development. The corpus has been affiliated with Universal Dependencies (UD) project as UD Romanian-Nonstandard Dependency Treebank; it contains 26225 sentences, 572436 tokens, part of them being collected and annotated in the Republic of Moldova. The paper concentrates on the corpus parsing using MaltParser tool.

Keywords: text corpus, treebank, dependency grammar, non-standard language, syntactic parser, automate parsing evaluation.

1 Introduction

Text corpora are the first and most essential linguistic resources in the Natural Language Processing (NLP) field of research. Corpus value is determined by its annotation, the additional information that can be as diverse as the corpora themselves. The advantages of corpus annotation are multi-functionality, re-usability, and easy of exploitation both by humans and computers. Morphological information is considered as a basic one. As a rule, the next level of annotation is syntactic one which indicates syntactic relations among words in sentences. These relations form a graph in the shape of a tree and such annotated corpus is referred to as a treebank.

The corpus presented in the paper is the current version of Alexandru Ioan Cuza University (UAIC) Romanian Dependency Treebank (UAIC-RoDia-DepTb). It has obtained the International Standard

Language Resource Number (ISLRN 156-635-615-024-0)1 and it is not merely syntactically annotated but it is presented in several formats: (1) the syntactic classic, containing more than 39,000 sentences, (2) the UD syntactic, and (3) a new syntactic-semantic one. Moreover, it is the biggest syntactically annotated Romanian corpus.

2 Related Work

Several corpora have been created for the Romanian language. CoRoLa [1] is a representative corpus of Contemporary Standard Romanian. Now, it contains more than 1,200 million words in written form and about 300 hours of oral texts with associated transcripts. It contains only Contemporary Standard Romanian and has only morphological annotation. Our corpus is annotated syntactically; it increases the corpus value.

Currently, the international community increased its interest in preserving and digitizing the cultural heritage, i.e. in the processing of old texts. In Romania as well as in Moldova, multiple old Romanian texts have been written using a specific Old Romanian Cyrillic alphabet. Texts with Cyrillic letters have been scanned and recognized using an Optical Character Recognizer (OCR) adapted and trained for these specific texts at the Institute of Mathematics and Computer Science in Chisinau [2]; then they have been transliterated using a special tool for the various versions of Old Romanian Cyrillic in the modern Latin. Our corpus contains the texts obtained by transliteration.

UAIC-RoDia-DepTb contains several parts with folklore. The Folk corpus of verses in Romania and Republic of Moldova is unique by combining folk texts and dependency annotation [3].

3 Corpus Annotation Format

The Romanian Non-standard UD corpus has been developed starting from the Alexandru Ioan Cuza University Romanian Diachronic Treebank (UAIC-RoDia). At present, the UAIC Dependency Treebank is the largest one for Romanian.

RoDia (Romanian Diachronic) corpus contains non-standard types of texts [4]. The standard language is rarely used in human communication; in our corpus we concentrate on the annotation of non-standard text types such as oral regional fiction, social media communication, poetry, Old Romanian texts and others.

The whole corpus has morphological and syntactic annotation using the dependency grammar conventions. There are 544 labels used for the complete morphological analysis in our corpus, part of the labels used in the MULTEXT EAST project [5], and 44 syntactic labels are used for the dependency relations.

The next step of our corpus development was its affiliation with Universal Dependencies (UD) [6]. UD annotation is coded in CoNLLU format[1]. It presents text files in which each line contains one word of the text with all its description separated by tabulation. The UD annotation convention highlights words with full meaning and the relational words are subordinated to them. In the UAIC convention, the relational words are the heads for the dependent words. In the UD system, it is easier to compare texts in very different languages and to emphasize the relation between the morphology and the syntax. UD annotation uses 17 very general PoS tags and 37 dependency labels.

The current numbers of sentences in both formats described above are shown in Table 1.

Table 1. The Current Statistics of Two Formats of UAIC Treebank

Nr.	Format	Sentences	Tokens
1	UAIC Syntactic (basic)	39,575	836,196
2	UD Syntactic	21,403	449,959

4 Parsing Evaluation

There are two Romanian UD corpora: RRT which contains modern Romanian texts and Non-Standard Corpus with old texts and folklore.

[1] http://universaldependencies.org/format.html

By May 2020 the volumes of these corpora were: 218,511 tokens total in RRT and 449,959 tokens in Non-Standard. Statistics of morphology tags as well as some other statistics for both corpora is presented on the UD page[2].

The main aim of our work is the creation of the gold standard corpus to be used for future training of part of speech taggers and syntactic parsers; its volume should be enough for reliable parsing with minimum errors. We tested MaltParser[3] which is considered the basic one for UD corpora on our corpus. We evaluated nine parsing algorithms of the parser[4] training on train part of our UD corpus and testing on the testing part. The results are presented in the Table 2.

Table 2. The results of the Malt-parser evaluation on our UD corpus

Nr.	Model	UAS	LAS
1	planar	82.28	74.36
2	stackeager	82.54	74.48
3	2planar	82.61	74.82
4	nivreeager	82.81	74.82
5	stacklazy	82.99	74.94
6	stackproj	82.99	74.94
7	nivrestandard	82.91	74.97
8	covnonproj	83.10	75.43
9	covproj	83.39	75.68

We measured the accuracy using two most commonly used metrics for dependency parsers: the labeled attachment score (LAS) and the unlabeled attachment score (UAS) on the word level [7]. LAS is calculated as the percentage of correct links with correct labels of all links in the test corpus; UAS is calculated as the percentage of correct links

[2] http://universaldependencies.org/treebanks/ro-comparison.html

[3] http://maltparser.org

[4] http://maltparser.org/guides/opt/quick-opt.pdf

ignoring their labels.

Our previous (autumn 2018) version of Non-Standard corpus resulted in LAS = 72.45 and UAS = 80.73. Since then, the train part of the corpus grew from 7965 to 10144 sentences. We experimented with different training part sizes; the learning curve is shown in Figure 1. The figure shows that the accuracy growing speed is slowing down while the corpus reaches 3000 sentences. It means that further gain in accuracy will cost us more and more additional text.

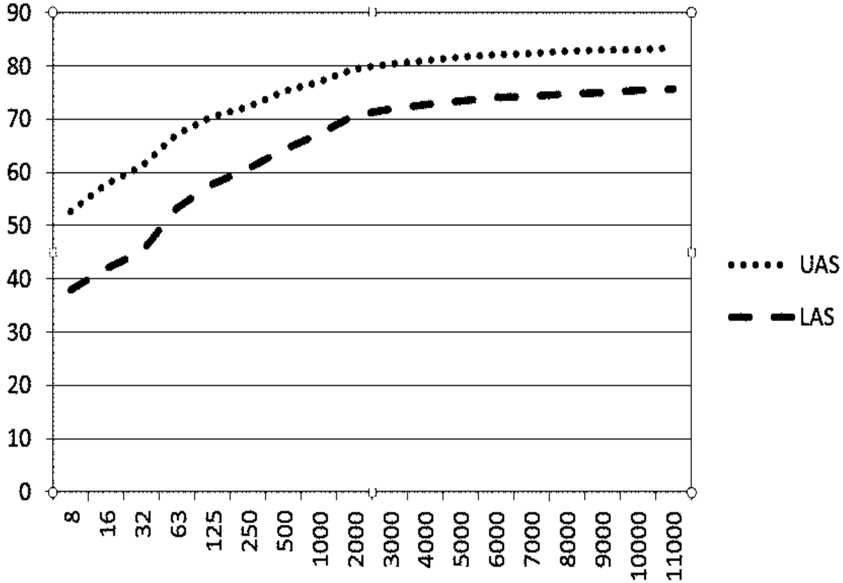

Figure 1. Dependency between the training corpus size and parsing accuracy

Tables 3 and 4 present more details of the the parsing errors. Table 3 contains two columns: Part of Speech (PoS) accuracy; syntactic link label accuracy. Table 4 presents triple's (PoS - link label - PoS) accuracy. There are 41 syntactic link labels for Romanian UD corpora, including several specific labels such as labels for the Romanian reflexive pronouns and its types written as “expl”: expl:pv, expl:poss, expl:impers. expl:pass. We also collected statistics for 253 types of so-called triples to see which pairs of PoS are connected by which links; Table 4 contains triples with the best and worst accuracy.

Table 3. The best and the worst parsing accuracy for various parts of speech and link types

PoS	Acc.	Link	Acc.
ADP	0.96	det	0.97
DET	0.94	aux	0.97
AUX	0.89	case	0.97
PART	0.88	expl:pv	0.94
...	...	...	...
ADV	0.73	expl:poss	0.19
INTJ	0.72	expl	0.18
NOUN	0.64	advcl:tcl	0.08
VERB	0.59	orphan	0.04
X	0.16	expl:pass	0.03

The first column of Table 3 presents accuracy for various parts of speech and it is not exactly correlated with their frequency. For example, NOUN and VERB, one of the most frequent PoS have almost the worst accuracy; it can be explained by the multitude of relations these PoS have. The second column presents link types and it is seen that it actually correlates with the first column with PoS. The triples presented in Table 4 explain some correlations from the previous table. For example, **ADP** and **case** appear in **ADP case PRON** triple; **DET** and **det** appear in **DET det PROPN** and **DET det PRON**, all these triples having accuracy 1.

Table 4. The best and the worst parsing accuracy for various triples: parts of speech - link type - part of speech

Triple	Accuracy
ADP case PRON	1
DET det PROPN	1
DET det ADJ	1
ADP case NUM	1
...	...
NOUN orphan NOUN	0.1
VERB advcl:tcl VERB	0.09
NOUN nsubj:pass VERB	0.09
PROPN nsubj:pass VERB	0.08
PRON expl:pass VERB	0.02

5 Conclusion

The paper presents an ongoing work on the development of the corpus of non-standard texts with syntactic annotation. Several efforts have been made to enrich the corpora with various more difficult for the annotation examples such as folklore, chat and Old Romanian. The annotated and manually corrected texts would serve as a training corpus for the statistical parser. Our current goal is to increase the volume of Old Romanian and folklore parts of our corpus until we get the acceptable accuracy of automate parsing for these types of texts. Our affiliation to UD increased the visibility of our common efforts and opened the perspectives of participating in the international projects.

References

[1] V. Barbu Mititelu, E. Irimia, and D. Tufiş. *CoRoLa – The Reference Corpus of Contemporary Romanian Language.* Proceedings

of the Ninth International Conference on Language Resources and Evaluation. (LREC-2014), pp. 1235-1239, 2014.

[2] S. Cojocaru, A. Colesnicov, and L. Malahov. *Digitization of Old Romanian Texts Printed in the Cyrillic Script.* Proceedings of DATeCH 2017, pp. 143-148, 2017.

[3] V. Bobicev, T. Bumbu, V. Lazu, V. Maxim, and D. Istrati. *Folk poetry for computers: Moldovan Codri's ballads parsing.* Proceedings of the 12th International Conference Linguistic Resources and Tools for Processing the Romanian Language, CONSILR-2016 pp. 39-50, 2016.

[4] C. Mărănduc, C.-A. Perez, L. Malahov, and A. Colesnicov. *A Diachronic Corpus for Romanian (RoDia).* Annual Conference of the German Linguistic Society, p.151, 2017.

[5] Tomaž Erjavec. *MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages.* Language Resources and Evaluation, 46/1, pp. 131-142, 2012.

[6] J. Nivre. *Towards a Universal Grammar for Natural Language Processing.* In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9041. Springer, Cham., 2015.

[7] J. Nivre and C.-T. Fang. *Universal Dependency Evaluation.* Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies. (UDW 2017), pp. 86-95, 2017.

Victoria Bobicev[1], Cătălina Mărănduc[2]

[1]Faculty of Computer Science,"Al. I. Cuza" University, Iasi, Romania
E–mail: `catalinamaranduc@gmail.com`

[2]Tehnical University of Moldova
E–mail: `victoria.bobicev@ia.utm.md`