TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: "TAIGA" SYNTAX TREE CORPUS AND PARSER

Abstract. The "Taiga" project unites the corpus and the syntactic parser, being created in a new field of the corpus linguistics: the material obtained primarily meets the needs of machine learning, rather than linguistic search. The authors consider in detail the methodology for constructing the corpus, balance, volume and composition of its' segments, format and quality of tagging – which meets the current requirements for the development of tools for processing Russian language. Within the framework of the project, the creation of a large and open-source syntactic corpus in the Universal dependencies format is planned.
Keywords. Corpus construction, web corpus, syntax parsing, machine learning, corpus representativity, parsers for Russian.

1. Introduction

Modern corpus linguistics relies heavily on the concept of web as corpus [Kilgarriff 2001], which regards the web as an inexhaustible source of linguistic and extralinguistic data in a vast number of languages. Text data on the Internet is already presented in electronic form and is available for downloading and searching through search engines. Under such favorable conditions, almost everyone is able to assemble a web corpus for their work, and the number of such big corpora (billions of words) grows every year: this is the WaC family [Baroni et al., 2009], Aranea [Benko 2014], the TenTen corpora [Jakubicek et al., 2013], and for the Russian language — the General Internet Corpus of Russian language [Belikov et al., 2013a]. At the same time, as noted in the work [Sharoff 2006], the newest web-corpora are often opposed to the national ones, which are expected to be more representative, or, according to [Sichinava 2002], "culturally representative". Texts for national corpora are selected by semi-automatic methods, with manual correction, and therefore, such data collections are expensive to build and still they exist only for a small number of languages. Thus, against the background of the growing volume of data, the volume of useful and accurately collected data remains small and expensive to receive. Therefore, there is a shortage of linguistic resources available for downloading and changing, which is more felt by developers rather than by lexicographers and linguists. But big, and at the same time, pure data in open-source is a guarantee of creating good NLP-tools for the language.

2. Actual corpus production principles

Annotated corpora today are needed not only for linguists, but for machine learners as well. For Russian language there already exists a number of corpus resources (the subcorpus with resolved homonymy of RNC, Open corpora of Russian, the subcorpus with the resolved homonymy of GICR, Syntagrus, etc.), but each of them has some hidden disadvantages. The main criteria that are relevant from the point of view of the engineering approach to linguistic development are the following:

1) open source or Creative Commons licence — the ability to use the material at one's discretion, to modify it, to add new data to it, and publish the re-ranked data is very important in the development process.
2) sufficient volume of the data (more than 10 million words) — thus it is possible to collect implicit information, for example about rare words and their compatibility, syntactic behavior of individual words and different meanings of homonyms.
3) minimum share of errors embedded in the data itself — this includes both the sufficient quality of linguistic markup, and the completeness of meta-text information, the ability to uniquely separate one segment or genre from another, to find out their balance.
4) representativeness — for each development task the data should be sufficient and it should represent all possible variability in unbiased proportions.
5) solvability in a given metric — adequacy of data composition and its' features to the task.

Today developers still have to collect the data meeting these criteria themselves, doing the same chain of downloading and tagging the texts from the Internet, having different degrees of understanding of linguistic data preprocessing.

3. Modern web-corpora for machine learning

As noted, the cheapest way to collect a voluminous training text collection is to resort to Internet resources. For Russian language, large collections of web corpora are assembled — projects like RuTenTen and Aranea Russicum, which are not available for downloading, unlike resources based on Common Crawl, but all these corpora are crawled from an unbalanced set of links and represent the "black box" statistically. The situation is slightly

better with GICR [Lagutin et al. 2016], however, only 2 million words are available for download.

Within our project, we offer a new corpus that meets the aforementioned development requirements. Taiga Corpus is available for download and modification under the Creative Commons license (CC BY 4.0)[1]. The corpus is assembled from open sources, which are selected for covering the main branches of NLP development: training of syntactic and morphological parsers, extraction of named entities, studying of readability, automatic text attribution, genre definition, thematic modeling, chat-bots training, sentiment analysis, etc. The volume of the corpus in beta-version is 50 million words, in the full version — 500 million words. This size for a balanced corpus has all the potential to show results equival to those for a larger but less accurate one (as showed, for instance, in [Kutuzov, Andreev 2015]). The texts are enriched with metatextual information, which is important for specific development tasks (more in Section 5). Annotation is fully automatic and is achieved using the combination of parsers: by now we have concerned MALT-Parser, SyntaxNet, UDPipe and some others, having examined their quality on literary data. Therefore, the corpus will be even more convenient for machine learning needs as being already applied to some relevant algorithms and accordingly cleaned.

4. Format

The corpus is stored in xml format in UTF-8 encoding with all the relevant metainformation tags. For each text, indent and paragraph structure is kept as in source. All the texts from each source separately have gone deduplication by URL, and are also filtered for non-UTF symbols, html-tags, non-breaking space, etc. by the BeautifulSoup[2] Python package.

5. Segments and features

By now, our corpus contains data from 8 resources, all documenting normative and modern Russian language (we considered "modern Russian" as a language of speakers younger than 60 years old). For training, we have chosen 3 main segments, all differing stylistically and by word frequency: literary texts, news and "other" — various resources for specific needs of NLP. Literary texts come from Russian Magazine Hall, and have such metatextual features as date, title, author, URL; news are taken from Lenta.ru, Interfax

[1] https://creativecommons.org/licenses/by/4.0/

[2] https://en.wikipedia.org/wiki/Beautiful_Soup_(HTML_parser)

and Komsomolskaya Pravda, and have as metaproperties date, title, rubrics, thematic tags, named entity tags, URL, sometimes author name; "other" is extracted from Stihi.ru and Proza.ru (only recommended authors), NPlus1, TV Subs. As literary texts and news are good for parser training, news and Stihi.ru, Proza.ru are also convenient for thematic modelling, news have also main named entity tags to train extraction, NPlus1 has expert-annotated text "difficulty", which can be used for readability studies, and TV Subs contain a lot of dialogues from films and are good for chat-bot training. Table 1 represents the distribution of main text parameters in our corpora:

Table 1

resourse	words	texts	authors	mean text length	rubrics
Stihi Ru	750 217	4167	107	177	34
Proza Ru	20 513 805	7527	82	2729	36
NPlus1	1 692 326	7696	34	221	26
Interfax	6 579 301	48 107	0	137	8
Koms.Pravda	5 000 341	45 503	652	109	986
Lenta Ru	7 001 491	34 399	0	198	38
TV Subs	28 403 842	3965	0	7163	0
Magazine Hall	216 763 813	47 629	0	4551	346

By now the corpora size is about 290 millions of words. The main ratios of the segments are presented in Diagram 1:

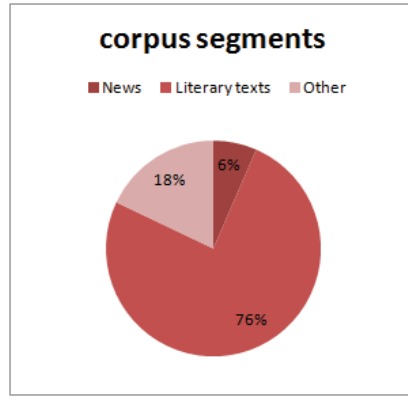

Diagram 1

Some of the data is distributed not quite normally, as you can see on Diagram 2 — yet we are going to fix such metaattribute problems when collecting the smaller version of 50 mln words.

Diagram 2 shows the distribution of texts by their difficulty in NPlus1:

Diagram 2

6. Future work

As we believe, morphological and syntactic standard for a big corpora for machine learning should be 1) concise 2) widely-adopted and compatible with international formats 3) suitable for rapid, consistent annotation by a human annotator 4) suitable for computer parsing with high accuracy 5) must be easily comprehended and used by a non-linguist (the last three points are Manning's Laws [Nivre 2016])). A new standard for multilingual morphological and syntactic tagging — Universal Dependencies[3] (UD) meets all the mentioned requirements. UD initiative has already developed treebanks for 40+ languages with cross-linguistically consistent annotation and recoverability of the original raw texts and now seeks to become the main annotation paradigm for many languages and the main evaluation tracks using it[4].

[3] http://universaldependencies.org/

[4] http://universaldependencies.org/conll17/

Our goal is to develop an open-source dependency parser (Taiga parser) and to obtain our corpus data tagged morphologically and syntactically in UD 2.0 format. We hope that our work will be useful for Russian natural language processing and will help developing new tools and projects.

7. Acknowledgments

The authors are sincerely grateful to Olga Lyashevskaya, Danil Skorinkin and Anastasia Bonch-Osmolovskaya, who expressed their deep insight and helpful advice at every stage of our work.

References

1. *Kilgarriff, A*. (2001), The Web as corpus. Proceedings of Corpus Linguistics.
2. *Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E.* (2009), The WaCky wide Web: A collection of very large linguistically processed Web-crawled corpora. In: Language Resources and Evaluation, 43, pp. 209–226.
3. *Benko, Vladimir* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: Petr Sojka, Ales Horak, Ivan Kopecek and Karel Pala (Eds.), Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655.Springer International Publishing Switzerland, 2014. pp. 257–264. ISBN: 978–3–319–10815–5 (Print), 978–3–319–10816–2 (Online).
4. *Jakubicek, M., A. Kilgarriff, V. Kovar, P. Rychly, and V. Suchomel* (2013), The TenTen corpus family. Lancaster: In: Proc. Int. Conf. on Corpus Linguistics.
5. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation. In: Web as Corpus Workshop (WAC-8).
6. *Sharoff, S.* (2006), Creating general-purpose corpora using automated search engine queries. In: Baroni and Bernardini (2006), pp. 63–98.
7. *Sichinava D.* (2002), K zadache sozdaniya korpusov russkogo yazyka [To the task of creation of Russian corpora]. In: NTI, cep. 2, 2002, no. 12.
8. *Benko V., Zakharov V. P.* (2016), Very Large Russian Corpora: New Opportunities and New Challenges. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016.
9. *Kutuzov A, Andreev I.* (2015), Texts in, meaning out: neural language models in semantic similarity task for Russian. In: Proceedeings of the international conference "Dialogue 2015" Moscow, May 27–30, 2015.
10. *Лагутин М. Б., Куратов Ю., Копылов Н.* (2016), Статистическая обработка результатов поиска в дифференциальных корпусах. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016" Moscow, June 1–4, 2016.
11. *Granovsky D., Bocharov V., Bichineva S.* (2010), Otkrytyi korpus: principy raboty i perspektivy [OpenCorpora: principles of work and perspectives].

12. *Nivre J.* (2016), Reflections on Universal Dependencies. Uppsala University. Department of Linguistics and Philology.

Tatiana Shavrina
NRU HSE, GICR
E-mail: rybolos@gmail.com

Olga Shapovalova
NRU HSE
E-mail: olya_shapovalova@bk.ru