Evaluation for morphologically rich language: Russian NLP

S. Toldova[1], O. Lyashevskaya[1], A. Bonch-Osmolovskaya[1] and M. Ionov[2]
[1]School of Linguistics, National Research University "Higher School of Economics", Moscow, Russia
[2]Department of Theoretical and Applied Linguistics, Moscow State University, Moscow, Russia

Abstract - *RU-EVAL is a biennial event organized in order to estimate the state of the art in Russian NLP resources, methods and toolkits and to compare various methods and principles implemented for Russian. Russian could be treated as an under-resourced language due to the lack of free distributable gold standard corpora for different NLP tasks* (*each team tried to work out their own standards*). *Thus, our goal was to work out the uniform basis for comparison of systems based on different theoretical and engineering approaches, to build evaluation resources, to provide a flexible system of evaluation in order to differentiate between non-acceptable and linguistically "admissible" errors. The paper reports on three events devoted to morphological tagging, dependency parsing and anaphora resolution, respectively.*

Keywords: NLP resources, evaluation, Russian, morphological taggers, dependency parsing, anaphora resolution.

1.Introduction

The NLP evaluation forum RU–EVAL started in 2010. A strong motivation for initiating the event was the need to independently review the stateoftheart of pos-taggers, parsers, and other NLP modules for Russian. The evaluation campaign is open both to academic institutions and industrial companies, and its major objective is to assess the current trends in the field and to promote the development of NLP technologies. This paper presents three evaluation campaigns held in 2010, 2012 and 2014, their task sets and results.

Although Russian computational linguistics history started more than 50 years ago (cf. the first MT research in 1955, Bar-Hilel 1960), up to the beginning of the 21 century the NLP teams, both research groups which inherited the Soviet tradition and new commercial industry labs, had been working as isolated units, with no interaction among them. The evaluation of standard NLP tasks such as pos-tagging, lemmatization, parsing, etc. requires certain unified standards and principles of annotation. In this respect, Russian could be treated as an under-resourced language since there were no free distributable gold standard corpora for various tasks. The problem was that each team tried to work out their own standards. The disjoint development of Russian NLP teams led to the co-existence of many different theoretical and engineering approaches. This plurality in approaches, tagsets, annotation principles, etc. makes the task of the evaluation very difficult. The paper summarizes our experience in working out the uniform basis for comparing different systems, building gold-standard evaluation resources, as well as providing flexible protocols for evaluation. The comparison of different approaches as well as working out unified standards reveal the controversial issues related to the nature of language data that are hardly amenable to formalization. For this reason we differentiate between non-acceptable and linguistically "admissible" errors which is helpful for determining the bottlenecks in language modeling for different levels of language description.

The paper is structured as follows. The remainder of Section 1 reviews the aims of the evaluation events for Russian, the topics and participants of the RU-EVAL tracks, and resources created by these events. Then we discuss the events on morphological tagging (Section 2), dependency parsing (Section 3), and anaphoric and co-reference relations detection (Section 4). Section 6 concludes.

1.1. Background

In the last decades, a special attention has been given to the evaluation of NLP systems. One of the main aims for such events is to provide the general grounds for independent evaluation via suggesting shared tasks for various NLP modules. The influence of such events on the NLP technology progress is indispensable. Among them are CLEF [1], Morpho Challenge [2], GRACE (for French, see [3]), Senseval/Semeval [4], MaltEval (for parsing, see [5]), CoNLL (for parsing, Named Entities recognition, anaphora and coreference resolution, [6]), ACE program [7], MUC7 [8], ARE (Anaphora Resolution Exercise, see [9]), etc. The majority of tasks are based on the testing sets for English (training and testing collections annotated according to shared tasks conditions, evaluation metrics and evaluation scripts etc.), though there are regular events for some other languages such as French (EASY) and Italian (EVALITA, [10]) as well as minor languages involved into multilingual task events. While preparing the testing sets evaluation procedures and resources for RU-EVAL events we took into consideration the experience of corresponding events (EVALITA as the most close to our conditions). However, we had to modify the tasks and evaluation principles for we have no opportunity to rely upon resources for Russian language for they are not easily

accessible for the majority of NLP teams and are not recognized by the majority as standards.

1.2. Aims of the evaluation campaigns for Russian: state of the art

As it has been mentioned above there are a lot of teams working with Russian language both those that are working on NLP tasks for more than 20-30 years and just newly organized start-ups, scientific research groups and business companies as well as educational groups from High School institutions. The majority of such teams are working in disjoint modes. Many of these teams start new NLP modules for Russian from the ground up.

The disjoint mode of development has led to the high diversification of standards and annotation schemes used in different NLP tasks by different teams. Thus, for instance we count more than 1000 different labels for dependencies for seven participants. The overlap in syntax tagsets was three-four tags only.

When the RU-EVAL campaign started in 2010 there were no generally distributed test collections, which could serve as a basis for systems comparison. There were some available resources (e.g. Russian National Corpus) or such morphological taggers as AOT or MyStem[1]. However, there was a need in a gold standard collection that consider discrepancies in tagging principles and allow comparison of systems based on quite dissimilar theoretical assumptions.

Thus, the aims of RU-EVAL events are: to consolidate the isolated NLP teams dealing with Russian language; to suggest gold standard collection that could serve the basis for comparison; to suggest principles of annotation for different tasks in Russian; to enumerate the discrepancies in existing theoretical and practical approaches; to suggest an evaluation scheme; to measure the basic level for different NLP tasks for Russian.

1.3. RU-EVAL events

In this perspective, the aim of the RUEVAL initiative is to promote the development of language technologies for the Russian language, by providing a shared framework to evaluate different systems and approaches in a consistent manner.

The NLP Evaluation forum RU–EVAL started in 2010. The participants are both academic teams and industrial companies. Organized on a fully voluntary basis, RU-EVAL was aimed at systematically proposing standards for Russian starting with the lower levels of linguistic analysis. The first three events were devoted to morphological tagging (2010, see [11]), parsing (2012, see [12]), anaphora and coreference Resolution (2014, see [13]).

[1] It's worth mentioning the Open Corpora project that nowadays suggests an open collection for morphological tagging training and evaluation.

The first NLP Evaluation forum focused on morphological taggers (see http://ru-eval.ru), bringing together 15 participants from Moscow, Saint-Petersburg, Yekaterinburg, Ukraine, Belarus and UK. We had seven different tasks in total. These were four tasks for tagging without disambiguation (lemmas, pos-tags, full set of grammatical tags and a special track for rare words) and three tasks for tagging with disambiguation (lemmas, pos-tags, full sets). The number of participants for each track is shown in Table 1.

Table 1. Participants of RU-EVAL

Task	Number of participants	Year
Lemmatisation	13	2010
Pos-tagging	13	2010
Full gram. Set	12	2010
Rare words	8	2010
Lemmatisation (disambig)	7	2010
Pos-tagging (disambig)	7	2010
Dependency parcing	7	2012
Anaphora	7	2014
Coreference	3	2014

In 2011-2012, syntactic parsing technologies were evaluated. Eleven participants expressed their interest in participation, seven of them submitted their answers. The task was to submit dependency parsing results. Only relations (head-dependent were taken into consideration) irrespective of assigned relation labels. There were two tracks depending on text types: general text collection and the News subcorpus.

The last event (2014) was devoted to the tasks of anaphora and coreference resolution and had seven participants in total. All of them submitted the anaphora resolution results and three of them participated in the coreference resolution track.

All the events had the preliminary discussions of evaluation details (corpora, formats, standards) with prospective participants. During the «Dialog-2011» (International Conference on Computational Linguistics in Russia[2]), a meeting on problems of syntactic parsers evaluation with the leading experts in the field was organized. In 2013 the round-table on anaphora and coreference resolution took place.

The complete cycle of the forum starts with working out mark-up scheme of the Gold Standard via analyzing the international practice of similar evaluation events, testing annotation schemes for Russian provided by potential participants, evaluating the inter-annotator agreement on preliminary test sets and ending with the final paper preparation. Our forum has an educational

[2] http://www.dialog-21.ru/en/

component: the expert group includes students who plan to work in the field of computational linguistics [14]. The evaluation cycle serves as a basis for a course in computational linguistics. It is worth mentioning that students not only do the routine assessment procedure, but take part in creating the forum design.

To sum up, the RU-EVAL has brought together a considerable number of IT companies and academic groups that work on Russian, and made it possible to assess the state-of-the-art in the field (so far, mostly in Russia).

1.4. Gold standard resources and evaluation schemes

As results of three RU-EVAL campaigns three sets of resources were created.

For each campaign the testing corpora include texts of different genres. Corpora consisted of fiction, news, non-fiction (science, law etc.) and texts from social networks (5%). We have 1 million tokens test sets for each of events. The gold standard sets are tagged manually by two annotators, the discrepancies being discussed and checked by a supervisor.

For morphology tagging we have a gold standard set of 3316 tokens and a testing list of rare words. For parsing we used manually annotated collection of 600 sentence (16000 tokens). The anaphora/coreference gold standard corpus at present consists of 185 texts (97 texts were used as learning set), containing 199681 tokens. The text length was from 5 up to 100 sentences, the longest one being 170 sentences long. These texts include 2900 anaphoric chains with 14405 total elements were annotated.

The forum devoted to parsers (RU-EVAL'2012) suggested the evaluation of existing syntactic parsers. It brought about several significant outcomes such as creation of a manually tagged and assessed gold standard treebank (800 sentences, available freely from http://rtb.maimbava.net/res01/rtb.php), treebank with parallel (1 mln. tokens, annotated by four participants, available from http://rtb.maimbava.net/res01/rtb.php). Besides the variations in theoretical and practical decisions between existing parsers have been analyzed.

The coreference/anaphora gold standard texts were split into sentences, pos-tagged with TreeTagger for Russian (we used a TreeTagger-based ([15]) part-of-speech tagger, a lemmatizer based on CSTLemma [16]. The learning set corpora is available from http://ant1.maimbava.net/.

Below we discuss each NLP task in more details from the point of view of language specific features and theoretical traditions which influence the development of NLP for Russian.

2. Morphology tagging

2.1. Russian as a morphologically rich language

The controversial issues we faced while working out the evaluation routine for Russian could be explained primarily by the fact that Russian like other Slavic languages is a morphologically rich language with a rather free word order. It is well known that the morphological richness increases the complexity of tagging [15]. Russian has a considerably large morphological tagset (cf. more than 4 592 unique full tags reported in [15], from which the top-1000 tags have each more than 40 occurrences in the 6M corpus [17]; cf. also 829 simplified tags used in [18]). There are 6 to 12 forms for nouns, ca. 30 forms for adjectives and more than 160 synthetic forms for verbs, including adjective-like participles.

Besides that, Slavic languages are inflectional with high index of fusion [19]. A bulk of grammatical categories is encoded in one short affix and this leads to high index of potential homonymy of word-forms. The word order is not as helpful in this case as in English and other German languages, for it is rather free in Russian, see Section 3.

Taking into consideration the diversity of existing taggers (see Section 1.2) the preparation stage of the event included their comparison. While comparing various approaches we come across the following linguistically motivated issues that need special treatment.

The size of verb paradigm could vary depending on whether two stems are organized into one paradigm or not. Many categories differentiate morphosyntactic classes and, thus, pertain to a stem and not to a certain morpheme. For example, there is no regular affixation for expressing aspect (imperfective VS perfective) in verbs, so aspect is a characteristic of a stem. Thus, the systems vary in whether each stem is lemmatized as a separate item, or a pair of stems are assign to one lemma of a fixed aspect (for some systems it is imperfective variant and for others perfective variant). The lemmatization of the so-called reflexive verb pairs is also non-trivial for some of the lexemes have regular pairs and are united under one lemma while in other cases the semantic relation between a verb and its derivative with the reflexive affix *–sya* is dubious.

The principles of pos-assignment could differ depending on the theoretical assumptions whether a pos-tag is determined by a lexeme paradigm (a set of affixes) or a word syntactic position. Thus, there is a sufficient variation in pronouns pos-tagging. For this reason, we do not take this class into account in evaluation procedure. One more complicated issue as far as the pos-tagging standard is concerned is the differentiation of conversion cases and the regular forms with the corresponding syntactic function: e.g. participle vs. "verbal" adjectives, adverbs vs. predicatives etc. The latter oppositions present a problem even for manual tagging.

The main problem both for theoretical modeling and for the method choice (c.f. context-based methods such as HMM or rule-based methods based on morphological parsing) is the

need of optimal balance between context and word-structure criteria.

2.2. Gold standard annotation and evaluation principles

The majority of Russian systems use rule-based methods without disambiguation since they disambiguate POS-tags in parsing modules, if needed. For this reason, we have four tracks for tagging without disambiguation.

We use a simplified list of parts of speech: separate parts of speech for inflected words such as Noun, Verb, Adjective, two syntactically defined classes, namely, Conjunction and Preposition, and ADV that includes adverbs, particles and other non-inflected words. We did not take into consideration pronouns tagging. We reduced the number of problematic stem-based tags such as aspect and voice for verbs (excluding voice for participles) and we do not take an opposition short vs. brief form for adjectives.

We had the only one tag for a token in gold standard. A system response was considered as true positive in case any of those given for the token matched the gold standard tag. We use accuracy as an evaluation measure.

2.3. Rare words track

A special track on rare words presupposed analysis of tokens which are, with a high probability, not included in the grammatical dictionaries of the dictionary-based systems. We put sentences with rare words into the common test set but evaluate responses separately.

A special word set for testing was comprised of 75 words. They represented the following classes of rare words: (a) those referring to productive word-formation types and rhyming false stimuli, eg. *frendjata* 'little friends', lemma *frendjonok* vs. *arrabjata* 'arrabiata' and those invented by authors, eg. *uvazila, slipkix*; (b) complex words with a dictionary-based second part, eg. *ul'trazhensvennoj* 'ultra - feminine'; (c) simple stem words with standard inflection affixes, eg. *turbijona* 'tourbillon.Gen.sg'; (d) rare and substandard forms of declination, eg. *visju* 'I hang', *derevjannee* 'more wooden'; (e) abbreviations.

2.4. Results and discussion

The general level of lemmatization and pos-tagging was considerably high for tagging without disambiguation. The median for lemmas task was 96.5 (max 99.3, min 72.8), pos-tagging 97.1 (max 99.4, min 72.8) and full tagging 94.8 (max 97.3, min 31.9). There were predictably lower scores on the rare words track, the median being 62 with maximum 72 and 4 as minimum.

Figure 1 shows results of tagging with disambiguation, the median being 94.5 for lemmas and 95 for pos-tagging.

The analysis of typical mistakes and systems drawbacks revealed some more language features that presented

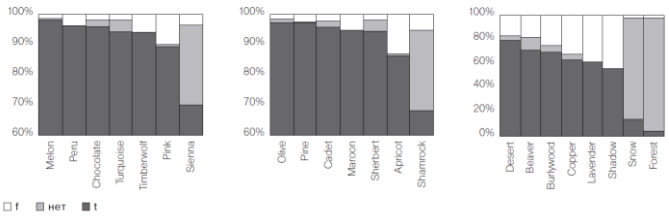

Figure 1

additional problems for highly inflected free-word order languages.

On the one hand, the rich inflectional system reduces the homonymy caused by conversion. On the other hand, it induces new homonymy types. Thus, there are many regular cases of grammatical forms homonymy such as genitive of masculine vs. nominative of feminine. Moreover, many systems generate potential forms that are never met in real texts (e.g. Russian preposition *dlya* 'for' is homonymous to a converb of the verb *dlit*' 'to prolong', that hardly could occur in real texts, it never occurs in National Russian Corpus). Consequently, the resulting tagging have a high coefficient of homonymy, e.g. 2.5-3 tags per token.

The majority of systems build up their analysis relying on inner word structure. The typical mistakes are the wrong analysis based on wrong detection of morpheme boundaries or matching it to a wrong morpheme annotation. Another class of multiple mistakes concerns a case when a token with clear morphological structure (e.g. adjectival suffixes) occupies a syntactic position of another part of speech (e.g. substantivized nouns, surnames with adjectival inflection system in Russian, etc.).

3. Dependency parsing

3.1. Challenges for the Russian parsing

The second forum took place in 2011-2012 and it was dedicated to syntax parsing.

Several properties of the Russian syntax make the syntactic parsing more difficult compared to English. The most important one is the free word order. In fact, word order (e.g. the order of major constituents such as subject, object) is mostly triggered by information flow (e.g. topic-focus hierarchy, prominence of participants in a profiled frame, emphasis, etc.). However, the order within individual constituents is more fixed, e.g. demonstratives and numerals usually precede nouns (but not always).

Subject is not obligatory in a finite clause in Russian, e.g. there are a lot of different types of impersonal constructions (c.f. [16] where adaptation of Universal set for dependency annotation to Slavic languages is discussed). A finite verb also could be omitted in a sentence (c.f. zero copula in the Present Tense). In this case it is unclear what could be chosen as a root for a syntactic tree. There are constructions in Russian, for example quantificational (numeric) groups, for which a controvertial evidence exists on what the relation direction should be. As a result some decisions concerning relation directions vary through systems.

3.2. Principles of syntax evaluation

Since different syntactic parsers employ different formalisms under the hood, it was very important to choose the right representation for the results. A preliminary study on the existing systems showed that most of them use dependency grammar representation. Depndencies were chosen as an output format. Participants who used representations other than dependency trees, were asked to convert their results. There was no unified treebank, every team had to start from scratch. Therefore, it was impossible to use a unified tag set, so we decided not to evaluate prediction of syntactic relations types. Only correct head detection for a node was evaluated.

Another important decision was to ignore some types of differences between the gold standard and markup provided by participants. The main assumption of an assessment procedure was that there is no 'correct' answer in some situations. Only divergences motivated neither by theoretical nor practical decisions were counted as mistakes.

The evaluation corpus consisted of untagged texts of various types: fiction, non-fiction, news and texts from social networks. Since systems performed different text preprocessing procedures, the corpus had been tokenized beforehand. A small part of the corpus (600 randomly selected sentences) was used as a gold standard for the assessment. It was manually tagged by two annotators independently.

Since verb-argument structure relations are mainly encoded by grammatical case and prepositions, the role of word order in the recognition of semantic-syntactic relations shrinks dramatically.

The results are presented in table Table 1

Table 2

System Name	P	R	F1
Compreno	0,952	0,983	0,967
ETAP–3	0,933	0,981	0,956
SyntAutom	0,895	0,980	0,935
SemSyn	0,889	0,947	0,917
Dictum	0,863	0,980	0,917
Semantic analyzer group	0,856	0,860	0,858
AotSoft	0,789	0,975	0,872

The best results have been achieved by the systems based on the manual rule-based approach. Both have a thoroughly elaborated ontologies and lexicographic resourses. However, low-time-consuming systems, such as SyntAutom, have also proved to be reliable. One of the systems, Russian Malt (presision 0,912), was based on the machine-learning technology. It used the SynTagRus Treebank as a learning corpus and achieved the third-highest results (the results are not shown in the chart since the system participated outside the competition).

This event has shown that although Russian is a free-word order language with a rich morphology, the quality of syntactic parsing is quite high. The majority of Russian parsers override the difficulties by developing semantic components and integrating statistical approaches into the rule-based systems.

4. Anaphora and coreference resolution

4.1. Anaphoric and coreference relations in Russian

Anaphora and coreference resolution event in 2014 started with the discussion of what types of relations we would like to detect.

Besides various general complications (annotation of appositive NPs like in *Petrov, the director of* …, annotation of abstract notions coreference) etc. Russian has some specific properties. It lacks the definiteness as a grammatical feature. A bare noun phrase without any determiner (demonstrative or a possessive pronoun) is a standard noun phrase (NP) type for non-first referent mention. There could be no clue in an NP for whether it is a newly introduced referent or before mentioned. For this reason, there are three nearly equal possibilities for such a NP interpretation: (a) a NP refers to one of the before mentioned referents (belonging to an existing coreference chain), (b) it introduces a new specific referent or (c) it is a generic NP. The differentiation of these three types is difficult even for human annotators. Another complication is due to free word order in Russian. It is not a rare event when a reflexive pronoun as *svoj* precedes its antecedent (as in *Svoji fotografii Petrov nikomu ne pokasyvajet* – lit.'[his own]$_1$ photos Petrov$_1$ to nobody shows'.

Thirdly, Russian is a so-cold pro-drop language, there are cases when a zero pronoun is used to refer to a subject of a clause, the omitted overt referent mention could influence the overestimation of distance between an NP and its coreferring NP from previous discourse.

There are also syntactic zeros for non-finite constructions. These are so called PROs, which are controlled by an NP from another clause and whose overt expression is ungrammatical. The relative frequency of the latter in Russian texts influence significantly on the coreference chains properties. A preliminary comparative research of coreference in Russian, Czech and English [20] has shown that the number of chains in Russian differs from those in English and Czech. According to the authors, the possible explanation is that the number of non-finite subordinate clauses such as infinitival or converb constructions have PRO in the subject position. Thus, the difference in coreference chaining could be strongly influenced by the clause structure of a sentence.

4.2. Gold standard annotation and evaluation principles

The campaign was devoted to both coreference chains extraction and anaphora resolution. The main aim was to track pronominal or all the mentioning of one and the same entity through the text. As previous events this event was the first

pilot run for Russian. The tasks were limited to the non-event anaphora; no implicit relations between corresponding NPs (such as part-whole, team-member etc.) were involved.

There were three participants in the first track and seven participants with total 17 runs for the anaphora resolution track. We prepared a little manually annotated training corpus consisting of nearly one hundred texts. Since each participating system has its own NLP pipeline, they used no predefined common standards for morphological and syntactic tagging learning set has no prerequisite morphological and syntactic annotation (including NP annotation).

In our annotation scheme, we addressed the identity relation between coreferential NPs. For there is no grammatical encoding for definiteness in Russian, special cases of distinguishing between discourse-new and discourse-old mentions as well as specific and generic reference were discussed in the annotation guidelines. We excluded from annotation procedure split antecedent cases, abstract notions, some classes of generic NPs. We took into consideration apposition and predicative NPs. However, the latter two NP types did not participate in evaluation. We took as markables NPs of maximal size. We mark potential semantic heads. There were participants who detected only heads as referring expressions. Moreover, the NP heads could vary through systems. Thus, in *admiral Pavel Nakhimov* the head could be *Pavel, Nahimov, admiral*. NPs matching evaluation was based on the head matching criterion in evaluation procedure.

The training data was distributed as a set of texts and a file with anaphoric chains information. In the anaphora dataset a chain consists of two elements: a pronoun (3rd person, possessive and reflexive, demonstratives and the relative pronoun kotoryj 'that'). In the coreference dataset a chain consists of all the NPs—mentions of the same entity with a set of attributes in the training corpus and without attributes in the testing set. While mapping system response NPs to Gold standard NPs we use soft criteria for NP boundary matching, that is a potential head matching principle. We used standard measures for anaphora resolution track. These are precision, recall and F-measure. We use MUC-score for coreference resolution.

4.3. Results and discussion

The forum has shown that there are competitive teams that develop high-level (discourse level) NLP components on a considerably high level (some systems manifest nearly 80% precision for anaphora resolution). However, the task of anaphora resolution is complicated for Russian due to free word order and the absence of overt markers of NP referential status. The absence of free semantic resource as WordNet and freely distributed syntactic parsers make the task more difficult for NLP start-ups and new small teams. The anaphora and coreference resolution tracks have shown the impact of high quality lower level linguistic analysis to the quality of discourse analysis tasks.

5. Conclusions

The RU-EVAL 2014 has brought together a number of IT companies and academic groups that work on Russian NLP tasks (pos-tagging, parsing, anaphora and coreference resolution), and made it possible to assess the state-of-the-art in the field (so far, mostly in Russia). The forum has shown that there are competitive teams that develop NLP components on a considerably high level. However, these tasks have some peculiarities and complications due to high inflectional and fusional properties of Russian language, its free word order and the absence of overt definiteness markers for NP. The absence of free semantic resources as WordNet and freely distributed syntactic parsers make the task even more difficult for newly organized NLP small teams. However, the event was the challenge for those teams that conduct the experiments on various machine-learning techniques.

The event has the following practical outcomes:

- the baselines for three NLP tasks were evaluated;
- the guidelines for tagging according to GS principles have been compiled and tested for Russian;
- new anaphora resolution systems for Russian arises at stretch due to the RU-EVAL 2014 campaign;
- the manually tagged standard sets for morphological tagging, parsing and anaphora and coreference resolution arises;
- new resourses for anaphora and coreference annotation (RuCor) are made available through http://gs-ant.compling.net/ and http://ant1.maimbava.net/ (the latter is to be moved to the former URL);
- RPTB - the Russian Treebank with parallel annotation of four systems (1 million tokens) is available at http://otipl.philol.msu.ru/~soiza/testsynt/
- a new Treebank for Russian (RTB) with manually annotated Subject, Object, Attributive modifier relations has come into being (http://rtb.maimbava.net/res01/rtb.php);
- the created corpora includes the wide variety of genres and various types of coreference relations.

The organizers hope that these corpora would be helpful for other NLP teams for the experiments on coreference resolution algorithms.

6. Acknowledgments

We acknowledge support from the Russian Foundation for Basic research (grant No. 15-07-09306). We thank the Lomonosov Moscow University students RU-EVAL team, and Dmitrij Gorshkov for software support, the RuCor and Treebanks interfaces creation.

7. References

[1] URL: http://www.clef-campaign.org/

[2] URL:http://research.ics.aalto.fi/events/morphochallenge/

[3] Gilles Adda et al. “The GRACE French Part-of-Speech Tagging Evaluation Task”. LREC, 1998.

[4] URL: http://www.senseval.org/

[5] Jens Nilsson, Joakim Nivre. “MaltEval: An Evaluation and Visualization Tool for Dependency Parsing”. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco, May. 2008.

[6] URL: http://ifarm.nl/signll/conll/

[7] URL: http://www.nist.gov/speech/tests/ace/

[8] Hirschmann L. “MUC-7 coreference task definition. Version 3.0”. Proceedings of the 7th Message Understanding Conference, 1997.

[9] Orasan C., Cristea D., Mitkov R., and Branco A. “Anaphora resolution excersice: an overview”. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco, May. 2008.

[10] Magnini, Bernardo, et al. “Evaluation of Natural Language Tools for Italian: EVALITA 2007”. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco, May. 2008.

[11] Lyashevskaja Olga, Astaf'eva Irina, Bonch-Osmolovskaja, Anastasia, Gareyshina Anastasia, Grishina Julia, D'jachkov Vadim, Ionov Maxim, Koroleva Anna, Kudrinskij Maxim, Lityagina Anna, Luchina Elena, Sidorova Evgenia, Toldova Svetlana, Savchuk Svetlana., Koval' Sergej. “Evaluation of the automated text analysis: POS-tagging for Russian”. (Ocenka metodov avtomaticheskogo analiza teksta: morfologicheskije parsery russkogo jazyka). Proceedings of the International Conference on Computational Linguistics Dialogue-2010. [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2010”], pp. 318-327.

[12] Gareyshina Anastasia, Ionov Maxim, Lyashevskaya, Olga, Privoznov Dmitry, Sokolova Elena, Toldova Svetlana. (2012). RU-EVAL-2012: Evaluating Dependency Parsers for Russian. Proceedings of COLING 2012: Posters. pp. 349-360. URL: http://www.aclweb.org/anthology/C12-2035.

[13] Toldova S., A. Roytberg, A. A. Ladygina, M. D. Vasilyeva, I. L. Azerkovich, M. Kurzukov, G. Sim, D. Gorshkov, A. Ivanova, A. Nedoluzhko, Y. Grishina. "RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian". Proceedings of the International Conference on Computational Linguistics Dialogue-2014. [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2014”], pp. 681695.

[14] Bonch-Osmolovskaya A., Lyashevskaya O., Toldova S. “Learning Computational Linguistics through NLP Evaluation Events: the experience of Russian evaluation initiative”. ACL 2013, Fourth Workshop on Teaching Natural Language Processing. Sofia, 2013, pp. 61-65.

[15] Dalal A. et al. “Building feature rich pos tagger for morphologically rich languages: Experience in Hindi”. ICON, 2007.

[16] Jongejan B., Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1* (ACL '09), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 145-153.

[17] Sharoff S., Nivre J. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. Computational linguistic and intellectual technologies, 2011, pp. 591-604.

[18] Sokirko A.V., Toldova S.Ju. “Sravnenie effektivnosti dvukh metodik sniatiia lexicheskoi i morfologicheskoi neodnoznachnosti dlia russkogo jazyka [The comparison of two methods for the morphological umbiguity resolution in Russian language]”. Internet-matematika. Moscow, 2005.

[19] Marszałek-Kowalewska K., Zaretskaya A., Souček M. Stanford Typed Dependencies: Slavic Languages Application. Advances in Natural Language Processing. 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17-19, 2014. Proceedings. Springer International Publishing. 2014, pp 151-163.

[20] Nedoluzhko A., Toldova S., Novak M. Coreference Chains in Czech, English and Russian: Preliminary Findings. Proceedings of the International Conference on Computational Linguistics Dialogue-2014. [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2015”].