RU-EVAL-2014: EVALUATING ANAPHORA AND COREFERENCE RESOLUTION FOR RUSSIAN

Toldova S. Ju.[1,2] (toldova@yandex.ru),
Roytberg A.[1, 3] (cvi@yandex.ru),
Ladygina A. A.[2] (aladygina@yahoo.com),
Vasilyeva M. D.[2] (linellea@yandex.ru),
Azerkovich I. L.[2] (iazerkovich@gmail.com),
Kurzukov M.[2] (mkurg@ya.ru),
Sim G.[2] (sim.ge@yandex.ru),
Gorshkov D. V.[2] (d.gorshkoff@gmail.com),
Ivanova A.[2] (ivanastas@gmail.com),
Nedoluzhko A.[4] (nedoluzhka@gmail.com),
Grishina Y.[5] (jul_gr@mail.ru)

[1] Natioanl Research University Higher School of Economics,
Faculty of Philology, Myasnitskaya 20, 101000 Moscow, Russia

[2] Moscow State University, Philological Faculty, Dept.
of Theoretical and Applied Linguistics, Leninskie gory, GSP-1,
119991 Moscow, Russia

[3] IMPB (Institut of mathematical problem in biology) RSS

[4] Charles University in Prague

[5] Applied Computational Linguistics, University of Potsdam

The paper reports on the recent forum RU-EVAL – a new initiative for evaluation of Russian NLP resources, methods and toolkits. The first two events were devoted to morphological and syntactic parsing correspondingly. The third event was devoted to anaphora and coreference resolution. Seven participating IT companies and academic institutions submitted their results for the anaphora resolution task and three of them presented the results of the coreference resolution task as well. The event was organized in order to estimate the state of the art for this NLP task in Russian and to compare various methods and principles implemented for Russian. We discuss the evaluation procedure. The anaphora and coreference tasks are specified in the present work. The phenomena taken into consideration are described. We also give a brief outlook of similar evaluation events whose experience we lay upon. In our work we formulate the training and Gold Standard corpora construction guidelines and present the measures used in evaluation.

Keywords: NLP evaluation, anaphora/coreference resolution, scoring metrics, coreference corpora

1. Introduction

The NLP Evaluation forum RU–EVAL started in 2010 as a new initiative aimed at independent evaluation of NLP systems for Russian. The Third evaluation campaign (2011–2012) focuses on anaphora and coreference resolution. The main objective of the Forum is to promote the development of language technologies for Russian. It unites separate teams dealing with NLP for Russian both from academic institutions and from industrial companies; provides the unified platform for the evaluation of technologies and algorithms; suggest the expertise for the current state-of-the-art in the field.

The organization of the forum is based on the experience of independent NLP systems evaluation events for different languages as well as multilingual evaluation events such as MUC (Message Understanding Conference), EVALITA (evaluation for NLP systems in Italian), ConLL and some others. This campaign is a pilot event for anaphora/coreference resolution tasks for Russian held for the first time. Thus, on the one hand we follow the basic principles of data design and evaluation worked out for afore mentioned events. On the other hand, we used a modified (simplified) conditions.

The forum also has an educational component: the expert group includes students and postgraduates in computational linguistics. It is a good opportunity for them to have a hands-on experience of how the NLP tools work.

The first NLP Evaluation forum focused on morphological taggers (see http://rueval.ru/news.html, [Lyashevskaya et al. 2010], bringing together 15 participants from Moscow, Saint-Petersburg, Yekaterinburg, Ukraine, Belarus and UK. In 2011–2012, syntactic parsing technologies were evaluated [Toldova et al. 2012]. Seven participants took part in the campaign. The results of four participants are available as parallel Treebank now (URL: http://otipl.philol.msu.ru/~soiza/testsynt/).

The present campaign is devoted to two tasks such as coreference chains extraction and anaphora resolution. The main aim of the tasks is to track pronominal or all the mentioning of one and the same entity through the text. The anaphora/coreference resolution modules are important components for the Information extraction systems (named entities recognition and fact extraction tasks in particular) as well as for the MT systems. These NLP components could improve the results for the text summarization and text classification tasks.

The aim of both anaphora and coreference resolution components of an NLP system is to find all the mentions in the text that refer to the same real-world entity, e.g. (1):

(1) *a) Probovali sravnivat' text* ***zapisnoj knizhki***$_x$ *s rukopisjami* ***Nahimova***$_y$ *<… >* *b) issledovatelej zainteresovala eshcho* ***odna jego***$_y$ ***zapisnaja knizhka***$_z$*,* ***kotoraja***$_z$ *hranitsya sejchas v sevastopol'skom musee. c) V otlichije ot* ***nashej***$_x$ ***eta knizhka***$_y$ (***knizhka***$_y$) *sohranilas' luchshe. d) Na oborote oblozhki* ∅$_y$ *rukoj* ***Nahimova***$_y$ *napisano:* ***Pavla Stepanovicha Nahimova***$_y$*. … e) Itak somnenij ne bylo.* ***Obe knizhki*** *prinadlezhali* ***admiraly Pavlu Stepanovichu Nakhimovuy****.*
lit. … (they) *tried to compare* (the) *text of* (the) *notebook*$_x$ *with* (the) *Nahimov's manuscripts…The researches were interested in one more* (his) *notebook*$_y$ *kept now in the Sevastopol museum. Unlike ours*$_x$ *this book*$_y$ (book$_y$) *was preserved in better conditions* (as compared to ours). *It is written*

in Nahimov's hand on the reverse cover of this book: …Thus there was no doubt that both of the books had belonged to Admiral Pavel Stepanovich Nahimov.

In (1) we have five Noun Phrases (NP) referring to the same entity Pavel Stepanovich Nakhimov. The first, the third and two others are proper names (the surname Nakhimov, the full name Pavel Stepanovich Nakhimov and the military rank plus full name correspondingly) while the second one is a possessive pronoun. We have also another chain of NPs referring to an entity: these are *eshcho odna jego*$_y$ *zapisnaja knizhka, eta knizhka, kotoraja*. Besides full noun phrases there are pronouns such as *jego* 'his', *kotoraja* 'that'. These pronouns have no meaning by themselves, their interpretation dependents on previous expressions in context—its antecedent. Thus we have two types of problems in referent tracking in a text. The first one is the task to gather all the mentions of a referent in a text, using semantic, syntactic and other properties of corresponding NPs. The other one is while seeing an anaphoric element with no semantic clue for its referent to find out what particular noun phrase mentioned in the previous context could serve as such a clue.

We have two corresponding tracks for the present campaign: the coreference resolution task and the anaphora resolution task. The first one presupposes the detection of all the entity mentions and hence gathering all the NPs referring to a particular entity into a chain. The second task was to detect an antecedent of a pronoun in the text and hence to enumerate all two-element chains <antecedent, pronoun>.

This is the first pilot run of the tracks for Russian. Thus, the tasks were limited to the non-event anaphora; no implicit relations between corresponding NPs (such as part-whole, team-member etc.) were involved (see section 3.3 for details).

Both tasks are much more complicated than previous ones (syntactic and morphological annotation). The mainstream technologies in this field presuppose the morphological and syntactic analysis as pre-processing stages. The machine learning techniques are widely used with these NLP tasks.

There were three participants in the first track and seven participants with total 17 runs for the anaphora resolution track. The participating systems vary in their final purposes. Some of them were just experimental systems whose goal was to test some particular anaphora resolution techniques. Other participants are the full-scaled NLP systems having the anaphora/coreference resolution module as its component. For each system has its own NLP pipeline and no generally accepted standards of morphological and syntactic annotation exists for Russian no prerequisite information concerning noun phrase structure or morphological properties was given to the participants. A little manually annotated training corpus consisting of nearly one hundred texts was suggested to the participants in order to give the opportunity to the teams to test various machine learning technique.

The overall procedure was organized as follows: participants received a text collection, processed it in their systems and sent the result back in a unified format. Standard metrics such as precision, recall and F-measure were computed for the anaphora resolution and three types of metrics (section 4) used in reference resolution were assessed by comparing the result against the manually tagged Gold Standard (GS). The expertise of the task output was performed semi-automatically with double manual check of dubious cases. The set of coreference chains included mostly the NPs referring to the real-life entities (specific referents).

The evaluation procedure has manifested that there are systems for Russian that have quite high precision for the anaphora resolution task though for many systems the recall is low. The coreference task is more complicated. The general problems are as follows.

For languages like Russian the standard methods elaborated for English are not enough due to such features as free word order and no overtly expressed definiteness of a noun phrase.

2. Related evaluation campaigns and datasets

During the organization we relied upon similar evaluation events: MUC-7 [Hirschmann 1997], EVALITA [Uryupina et al. 2011], ARE (Anaphora Resolution Exercise) [Orasan et al.2008], SemEval 2010. First, we need to identify anaphoric and coreferential types which were assessed in these projects. In this section we give brief overview of evaluation initiatives, possible task definitions, available data resources

2.1. MUC

One of the first evaluation events for the tasks under discussion took place within the Message Understanding Conferences (MUCs). The following definition for the coreference relations was suggested: “The coreference ‘layer’ links together multiple expressions designating a given entity”. Only links between noun phrases were considered. The main criteria for the task definition were the good interannotator agreement, the simplicity and speed of text annotation, the objective to create a corpus for independent research of coreference. The MUC project is the best-known example of coreference annotation, on which much subsequent work is based. The various systems results on the test sets from MUC-6 (1995) and MUC-7 (1998) coreference corpora are widely discussed in the literature (c.f. [Mitkov 1999], [Van Deemter, K., & Kibble, R. 2000] and others). The main principle for data annotation was the referential NP identity. We followed the basic MUC data annotation principles.

2.2. Anaphora Resolution Excersise 2007 (ARE)

The other influential event in the coreference resolution domain is Anaphora Resolution Exercise[1] that was held within the Discourse Anaphora Colloquium. There were four tracks depending on the anaphora types and the types of prior information. The data sets for the first two tracks were pre-annotated: the NPs were detected and tagged, some of them were tagged as NPs for which the detection of the antecedent is required. The data for two other tracks remained unannotated.

[1] Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2007)

2.3. SemEval 2010

The SemEval-2010 task is of special interest for us in the following respect: it is the task task on Coreference Resolution in Multiple Languages (six languages such as Catalan, Dutch, English, German, Italian and Spanish). Besides the well-studied languages, the under-resourced languages took part in the anaphora resolution procedure. Some of the participants used small training and testing data sets (e.g. 80 texts for Italian as a training set and 46 texts as a test set). The other issue of interest is that four different metrics were used in evaluation. The corpora were annotated with morphological, syntactic and partial semantic tags. One of the aims of the event was to learn out what was the impact of different levels of linguistic information.

2.4. Evalita 2011

In the latest anaphora resolution tasks such as Evalita-11 task the systems are required to recognize all the mentions of an entity and to cluster them into a chain irrespective of the NP type (including names, pronouns, zero pronouns, etc.) The singleton NPs are also included. Not only referring NPs are taken into consideration. All the NPs were marked and annotated [Uryupina et al. 2011]. The additional information on morphological and syntactic token properties was given as well as some mention and semantic types.

2.5. Some important corpora

Our data set includes training corpus and Gold Standard corpus that were designed considering the existing linguistic corpora for other European languages which contain coreferential annotations: OntoNotes [Hovy et al., 2006; Pradhan et al., 2007], Potsdam Commentary Corpus [Stede, 2004], ARRAU Corpus [Poesio & Artstein, 2008], Prague Dependency Treebank [Böhmová et al., 2003], VENEX Corpus [Poesio et al., 2004]. In our work we tried to rethink the existing methodology and implement it to Russian.

We also studied the annotation principles and annotation schemes used for coreference corpora presented in [Nedoluzhko et. Al 2009], [Nedoluzhko 2013], [Khudyakova 20011]. However these corpora are based on deep annotation schemes. The first one requires the syntactic analysis as the prerequisite condition. It also deals with implicit anaphoric relations such as bridging. The detailed multifactor annotation scheme is implemented for the latter corpus.

Thus our annotation scheme was based on MUC corpora principles. For the simplification of annotation task only identity relations were annotated, other types like bridging anaphora, predicative anaphora, coreference of events were excluded. We also decided to ignore these types of coreferential relations based on the results of several experiments to evaluate the capacity of existing Russian coreference resolvers.

Unlike SevEval 2010, we do not take into account expressions that occurred only once. We did not take into account morphological and syntactical tags (despite the fact that we included them into the output to make the manual evaluation easier).

3. Participants and data sets

3.1. Tasks

As it was mentioned above the tasks proposed in this exercise focused on problems related to anaphora and coreference resolution.

The purpose of the first task was to evaluate the quality of different pronoun resolution algorithms. The systems were expected to recognize pronouns which need to be resolved (we examined only personal, possessive, reflexive pronouns and a relative pronoun *kotoryj* 'which' as well) and find their antecedents. We use the lenient principle of evaluation assuming the transitivity of anaphoric relations. The antecedent NP established by a system should be a member of corresponding coreference chain in the Gold Standard set.

In the second task the participants were required to recognize referential expressions (proper names, nouns and pronouns, excluding zeroes) in an unannotated document and cluster them into coreference chains.

3.2. Participants

Eight NLP groups from Moscow and St. Petersburg expressed their interest in participating in the tracks of the Third RU-EVAL forum. They are: Compreno (Abby, Moscow), RCO (Russian Context Optimizer, Moscow), SemSyn K. Boyarsky, E. Kanevsky, St. Petersburg). Open Corpora (St.Petersburg), Mail.ru anaphora resolver (Mail.ru, Moscow), the system of Institute for Systems Analysis of Russian Academy of Sciences (ISA RAS), Sergej Ponomarev's system. Three teams took part in coreference resolution task as well. Some teams submitted results for different machine learning and rule based techniques. Thus, we have 17 answers from 7 systems for anaphora resolution track.

3.3. Main principles of dataset preparation

While preparing the data sets for the tracks we based on the principles of the MUC anaphora resolution corpora creation. The resulting corpus is the first open corpus annotated for coreference relations for Russian, and thus, the main purposes and principles of our corpus constructing are: the resulting corpora should be of open access in order to be freely distributed among the community;

1) the resulting corpora should be open access in order to be freely distributed among the community;

2) it should be reusable, open access, distributed in a machine readable format on the one hand and presented in human readable form on the other hand (the platform for easy annotation and visualization of data should be provided);
3) the corpus should include various genres (in contrast to the majority of coreference corpora);
4) the annotation procedure should be simple,
5) the high annotator agreement should be the crucial criteria for chain inclusion into the corpus;
6) only entity referring expressions are considered (no event anaphora);
7) only identity relation between NPs referents (expanded to near-identity in some cases) are under consideration (no bridging, split referents etc.).

3.4. Training and test corpora

This project required work with the widest possible selection of texts. For this reason, short texts or fragments of texts in a variety of genres have been included in the test corpus: news, scientific articles, blog posts and fiction. All texts were taken from publically available sources, such as Russian OpenCorpus, online library Lib.ru and Lenta.ru. For test corpus about 300 texts were used from each above mentioned source, to the total of 1342 texts. 20 texts of each section were used for annotating the Golden Standard.

In preparation for the Anaphora Resolution Event, a subcorpus of approximately 100 texts was selected as a training corpus for the participants. The texts were tokenized, split into sentences, pos-tagged with TreeTagger for Russian (we used a TreeTagger-based ([Schmid 1994]) part-of-speech tagger, a lemmatizer based on CSTLemma ([Jongejan, Dalianis 2009]) available at URL: http://corpus.leeds.ac.uk/mocky/).

These texts were used as a basis for training the participating systems, which later were tested against the whole corpus. The texts were from 5 up to 100 sentences long, the longest one being 170 sentences long. In total 2000 anaphoric pronoun—antecedent pairs and 1200 coreferential chains were annotated by hand. The texts were annotated by 2 annotators and the differences between their annotations were compared and analyzed. After that they were checked by supervisors who made the necessary updates.

The special Web-interface was designed by Dmitrij Gorshkov for corpus annotation. The tool uses MySQL database engine for corpus management. While constructing this tool we took into account some features of MMAX-2 mark-up tool (available at http://mmax2.sourceforge.net/, see [Müller C., Strube M. 2003]), Brat annotation tool (available at http://brat.nlplab.org/, see also [Nilsson Björkenstam, K., & Byström, E. 2012]) and some others. However we decided to use our own tool based on MySQL since it provided the flexible work with corpus data, convenient visualized Web-interface suitable for collaborative work on corpus annotation and the annotation comparison.

3.5. Gold Standard Preparation: basic principles and instructions

In the corpora (training and Gold Standard), the following NPs were annotated:

(1) pronominal and nominal NPs referring to real-world entities;

(2) non-specific (generic and abstract NPs) if they are antecedents of pronouns from our list (tagged for the anaphora resolution track only)

According to the results of several experiments on assessment of the capacity of existing Russian coreference resolvers, we decided to ignore:

1) bridging relations (c.f. part-whole relation in *zapisnaja knizhka* 'the note-book' and *oblozhka* 'the cover (of the note-book)' in example (1)),
2) discourse deixis (1st and 2nd person pronouns)
3) and coreference relations with a split antecedent in our annotation;
4) discontinuous expressions (c.f. "***Peter*** *came there with* ***Masha****. They…*") and split antecedents.

Development of our mark-up scheme was influenced by the annotation guidelines proposed in [Krasavina and Chiarcos, 2007]. These guidelines were created for annotating coreference in German and English. We used slightly modified annotation scheme and our own Guidelines (adapted for situation with Russian NLP systems).

Referential expressions subject to annotation are called markables (or groups in our corpus). Only NPs can be markables. Markables are maximal NPs excluding the relative clauses, the postpositional participle constructions and some other (see above). The appositive NPs are treated as separate markables (groups).

There is also a distinction between primary markables and secondary markables. Thus, following expressions should be annotated as primary markables:

1) 3rd person pronouns (1st person pronoun is marked if only it denotes the narrator in the text);
2) demonstrative pronouns;
3) reflexive pronouns;
4) possessive pronouns;
5) definite and possessive descriptions;
6) proper names and titles;

Some other markables, which are not annotated as primary markables, but are potential antecedents for them, are also considered to be secondary markables, e.g.:

7) indefinite descriptions if they are used for the first mentioning of an entity.
8) apposition NPs as *Admiral* in

(2) *Pavel Stepanovich Nakhimov, vydajuscshijsya* ***rossijskij*** ***<u>admiral</u>*—**
'Pavel Stepanovich Nakhimov, 'the great Russian Admiral';

9) predicative NPs as *Gallej* in

The comet was called ***<u>Gallej</u>*** *or as in He became a* ***<u>great</u>*** ***<u>scientist</u>****;*

10) first and second person pronouns in the direct speech constructions.

The main difference between primary and secondary markables is that the former are always annotated while the latter are annotated only if they potentially could serve as antecedents for some of primary markables. The secondary markables are do not participate in the evaluation score. However the system is not penalized for establishing coreference relations with them.

For each markable there is a number of attributes to be defined. These are, for example, functional type: 'def'—for the expressions referring to the entities, 'pred'—for the predicative referring expressions, 'appo'—for appositive expressions, 'ds'—for 1st and 2nd person pronouns in direct speech, 'misc' for some other dubious cases such as near-identity cases (c.f. *his book—this edition of the book*). We also use the tag 'meton' for metonymies for it helps to single out the cases that are highly difficult for the detection of coreference. We also use special set of attributes for NP structure: the separate attributes for different types of pronouns ('refl', 'dem', 'rel' etc.) and 'noun' for non-pronominal NPs.

The process of annotation is based on several principles, also described in the guidelines. These principles refer to the order of establishing coreferential links and forming coreferential chains as well as annotating markables of maximal size. For each group longer than one word we mark the potential semantic head. There were participants who detected only heads as referring expressions. Moreover the NP heads could vary through systems. Thus, in the example in (2) the head could be Pavel, Nahimov, admiral. The information about potential heads is used in the evaluation procedure (see below).

The data for training include: 1) the group (markable) offset presented as the shift from the beginning of a given text in symbols and the length of the fragment; 2) its ID, 3) the text ID; 4) the chain ID 5) the referential expression itself.

The training data was distributed as a set of plain texts and an xml-file with anaphoric chains information:

1) in the anaphora dataset a chain consists of two elements: a pronoun from a list of pronouns (3rd person, possessive 3rd person and reflexive, demonstratives and the relative pronoun *kotoryj* 'that');
2) 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.

The same format is used for the systems response set.

This information was used both for automated comparison with the Golden Standard and for manual check by the annotators.

4. Measures

Systems were examined in two tasks: anaphora resolution and coreference resolution. We use F-measure to combine recall and precision to evaluate pronoun resolution algorithms work. We use MUC, B^3 and CEAF to compute a more complicated recall and precision for coreference and then we use F-measure. The results of evaluation are published at http://rueval.compling.net/anaph/results.html.

Below we use GOLD for Gold Standard corpus chains (linked groups) and RESPONSE for the system response chains and groups. We use the principle of lenient groups matching (the same for both tracks): the boundaries of GOLD NP should intersect the boundaries of System NP. The GOLD NP head should be included into intersection. A System NP matches the only one GOLD NP. There is an exclusion from the rule: when System NP includes the GOLD NP and the head noun of the System NP is out of the GOLD NP there is no groups match. The inaccurate NP boundaries (as in *oblomok skaly v reku*—'the piece of the rock (fell)') are not penalized.

4.1. Measures for anaphora resolution (precision, recall, f-measure)

We used standard measures for anaphora resolution track. These are precision, recall and F-measure.

We use the "lenient" principle for assessment of matching the system response to the Gold Standard anaphoric pair. The true set of entities is a set of all pronouns enumerated in 3.5. in GOLD with their antecedents. We map the System chains to GOLD chains according to the following principle: a chain from GOLD that includes a particular pronoun corresponds to the chain in RESPONSE with the same pronoun.

We consider the system response True Positive if the chain in RESPONSE with a particular pronoun is a subset of GOLD chain with the same pronoun (lenient group matching principle).

Precision then shows what proportion of System pairs match the GOLD (what part of all true pronouns and true links was found). S is all pronouns with their link from system output and M is all true pronouns with true links from system output. So we can compute the precision:

(1) $P=M/S$.

Recall shows what part of system output is true pronouns with true links, to compute this we use the formula:

(2) $R=M/G$.

The formula for F-score is:

(3) $F=2PR/P+R$,

where F is F-score, R is recall and P is precision.

4.2. Measures for coreference resolution MUC, B-Cubed, CEAF

We use three measures for the coreference track evaluation: MUC, B^3 and CEAF.

4.2.1. MUC-score

MUC is a link-based metric (for MUC measures see e.g. [Chen & Ng, 2013], [Vilain et al.1995] etc.). Recall and precision are associated with links between the golden standard chains and the system chains.

Recall is computed as the number of common links between the golden chains and the system chains in a document divided by the number of links in the golden chains.

Precision is computed as the number of common links divided by the number of links in the system chains.

4.2.2. B-cubed

B-cubed is entity-oriented cross-document coreferencing measure ([Bagga and Baldwin, 1998]). B-cube is a more complicated harmonic mean of recall and precision.

To compute B^3 one needs to compute the recall_i and precision_i for each mention, and then to take an average of these pre-recall and pre-precision values to obtain the overall recall and precision.

Mention precision is a number of correct elements in the system output chain containing the mention divided by the number of elements in the system output chain containing the mention_i

Mention recall_i is a number of correct elements in the system output chain containing the mention divided by the number of elements in the golden standard chain containing the mention_i.

Hence, to compute overall recall and overall precision you need to average all mention recalls and mention precision respectively.

4.2.3. CEAF

In [Luo 2005] it was shown the main problem of B^3 algorithm: B^3 may use all chains more than once when computing recall and precision. Luo proposes 2 metrics which aligns entities in golden standard and system output. First of all CEAF requires the establishing of all one-to-one corresponding alignments between the chains in G(d) and the chains in S(d). CEAF uses the function *j* to compute the similarity between G_i and S_j where G_i and S_j are the chains from golden standard and system output respectively. Furthermore, the algorithm proposes the best alignment using the Kuhn-Munkres algorithm.

4.2.4. The Evaluation procedure

Taking the above described measures into account we made the revision of Gold Standard data. First of all we checked the consistency of annotators principles and "switched off" all the chains that seemed not referential or caused some problems and discussions between annotators. Then we implemented the procedure of groups mapping and chain mapping. A random sample of mapped groups were checked manually. A random sample of erroneous links as well as missed links was also manually checked. Then the corresponding scores were automatically calculated.

Conclusions

The RU-EVAL 2014 has brought together a number of IT companies and academic groups that work on Russian 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 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. 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 baseline for anaphora and coreference resolution for Russian was 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 set, consisting of nearly 200 texts annotated for anaphora and coreference chains is made available through http://gs-ant.compling.net/ and http://ant.compling.net/ (the latter is to be moved to the former URL);
- the created corpus includes a 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.

Acknowledgments

We would like to thank other MSU students who participate in corpora preparation: Darya Skorobogatova, Alexander Pechonyj, Viktoria Danilova, Alexander Kostyuk, Max Ionov. We also are grateful to them and to some evaluation event participants: Anna Sergeeva, Natalya Glazyrina, Ivan Hramoin for their help in training corpus annotation. We also would like to thank Olga Lyashevskaya, Anastasija Bonch-Osmolovskaya for their help in the event organization. We are also most grateful to the participants of the forum.

The work was partially funded by the LINDAT/CLARIN project (project LM2010013).

References

1. *Bagga, A., Baldwin B.* (1998). Algorithms for scoring coreference chains, Proceedings of the 1st International Conference on Language Resources and Evaluation, Granada, Spain, 28–30 May 1998, pp. 563–566

2. *Böhmová, A., Hajič, J., Hajičová, E., & Hladká, B.* (2003). The Prague dependency treebank. In Treebanks, pp. 103–127. Springer Netherlands.
3. *Chen Ch., Ng V.* (2013). Linguistically Aware Coreference Evaluation Metrics, Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP-13), 2013, available at: http://www.hlt.utdallas.edu/~vince/papers/ijcnlp13-coref.pdf.
4. *Dipper, S., Zinsmeister H.* (2012), Annotating Abstract Anaphora. Language Resources and Evaluation 46 (1), pp. 37–52.
5. *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.
6. *Hirschmann L.* (1997). MUC-7 coreference task definition. Version 3.0, Proceedings of the 7th Message Understanding Conference (1997).
7. *Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R.* (2006, June). OntoNotes: the 90% solution. In Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers, pp. 57–60. Association for Computational Linguistics.
8. *Jongejan, B., Dalianis, H.* (2009), Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike, 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. Suntec, Singapore : Association for Computational Linguistics, 2009. pp. 145–153.
9. *Khudyakova, M. V., Dobrov, G. B., Kibrik, A. A., & Loukachevitch, N. V.* (2011). Computational modeling of referential choice: Major and minor referential options, Proceedings of the CogSci 2011 Workshop on the Production of Referring Expressions. Boston (July 2011).
10. *Krasavina, O., Chiarcos Ch.* (2007) PoCoS: Potsdam coreference scheme. Proceedings of the Linguistic Annotation Workshop. Association for Computational Linguistics, 2007.
11. *Luo, X.* (2005, October). On coreference resolution performance metrics. In-Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 25–32). Association for Computational Linguistics. Available at: http://dl.acm.org/citation.cfm?id=1220579.
12. *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.* (2010). Evaluation of the automated text analysis: POS-tagging for Russian. [Morphological Ananlysis 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.
13. *Mitkov, R.* (1999). Anaphora resolution: the state of the art. School of Languages and European Studies, University of Wolverhampton., available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.6235&rep=rep1&type=pdf

14. *Müller C., Strube M.* (2003). Multi-level annotation in MMAX. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue.
15. *Nedoluzhko, A.* 2013 How Dependency Trees and Tectogrammatics Help Annotating coreference in Prague Dependency treebank. DepLing 2013, Prague. No. 44, ÚFAL, Charles University in Prague.
16. *Nedoluzhko, A., Mírovský, J., & Pajas, P.* (2009, August). The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague Dependency Treebank, Proceedings of the Third Linguistic Annotation Workshop, pp. 108–111. Association for Computational Linguistics.
17. *Nilsson Björkenstam, K., & Byström, E.* (2012). SUC-CORE: SUC 2.0 Annotated with NP Coreference. In Proceedings of the Fourth Swedish Language Technology Conference (SLTC), October 24–26, 2012, Lund.
18. *Orasan C., Cristea D., Mitkov R., and Branco A.* (2008) Anaphora resolution excersice: an overview, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May. 2008.
19. *Poesio, M., and Artstein R.* (2008) Anaphoric annotation in the ARRAU corpus, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May. 2008.
20. *Poesio, M., Delmonte, R., Bristot, A., Chiran, L., & Tonelli, S.* (2004). The VENEX corpus of anaphora and deixis in spoken and written Italian. University of Essex.
21. *Schmid, H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees, Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.
22. *Stede, Manfred* (2004) The Potsdam commentary corpus, Proceedings of the 2004 ACL Workshop on Discourse Annotation. Association for Computational Linguistics, 2004.
23. *Uryupina O., Poesio M.* (2011). Evalita 2011. Anaphora resolution task. In Proceedings of Evalita 2011.
24. *Van Deemter, K., & Kibble, R.* (2000). On coreferring: Coreference in MUC and related annotation schemes, Computational linguistics, 26(4), pp. 629–637.
25. *Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman* (1995). A model-theoretic coreference scoring scheme. In Proceeding of the 6th Message Understanding Conference (MUC-6), pp. 45–52. San Mateo, Cal.: Morgan Kaufmann.