Error analysis for anaphora resolution in Russian: new challenging issues for anaphora resolution task in a morphologically rich language

Svetlana Toldova **Ilya Azerkovich** **Anna Roytberg**
National Research University Higher School of Economics
21/4, Staraya Basmannaya Ulitsa, 105066, Moscow, Russia
toldova@yandex.ru, cvi@yandex.ru, iazerkovich@gmail.com

Alina Ladygina
Eberhard Karls Universität Tübingen
Geschwister-Scholl-Platz,
72074, Tübingen, Germany
aladygina@yahoo.com

Maria Vasilyeva
Lomonosov Moscow State University
1 Humanities Building, 1-51 Leninskie Gory,
GSP-1, 119991, Moscow, Russia
linellea@yandex.ru

Abstract

This paper presents a quantitative and qualitative error analysis of Russian anaphora resolvers which participated in the RU-EVAL event. Its aim is to identify and characterize a set of challenging errors common to state-of-the-art systems dealing with Russian. We examined three types of pronouns: 3rd person pronouns, reflexive and relative pronouns. The investigation has shown that a high level of grammatical ambiguity, specific features of reflexive pronouns, free word order and special cases of non-referential pronouns in Russian impact the quality of anaphora resolution systems. Error analysis reveals some specific features of anaphora resolution for morphologically rich and free word order languages with a lack of gold standard resources.

1 Introduction

Anaphora resolution, or the task of identifying noun-phrase antecedents of pronouns and adjectival anaphors in a text, is an essential step in the text-processing pipeline of NLP. Still, building an anaphora resolution module is challenging for text-mining systems, as it requires a high level of morphological and syntactic analysis at the first stages of the NLP pipeline. Nevertheless, this task has a long history of development and evaluation (e.g. the MUC-6 conference in 1995), and different aspects of anaphora resolution are well studied and have rich resource support, especially for English. However, Russian (as well as other Slavic languages) poses additional challenges for anaphora resolution, in particular, it has rich morphology, free word order and lacks articles. Furthermore, Russian is a relatively low-resourced language (Toldova et al., 2015) due to the lack of freely distributable gold standard corpora for different NLP tasks.

In our paper, we analyze the performance of Russian anaphora resolvers which participated in the RU-EVAL-2014 evaluation campaign (Toldova et al., 2014). RU-EVAL-2014 was dedicated both to anaphora and coreference resolution, but our study focuses only on anaphora resolution, as there were more participants in this task and the results obtained were more reliable.

The aim of this paper is to present quantitative and qualitative error analysis for different pronoun types (reflexives, 3rd person pronouns and relative pronouns). We identify and characterize a set of challenging errors common to state-of-the-art systems dealing with Russian. Error analysis enables us to compare the efficiency of different NLP approaches and detect errors that occur either due to language-specific issues or system defects that could be fixed.

In Section 2, we discuss the previous experience of anaphora and coreference resolution error analysis that we took into account. In the section 3, we give a short overview of RU-EVAL-2014, describe the data used for evaluation (RuCoref corpus) and the annotation scheme. Then, we briefly describe systems that took part in RU-EVAL-2014, the evaluation principles and systems' general performance. The qualitative and quantitative error analysis presented in section 4 reveals language specific features influencing system performance such as particular types of morphological ambiguity, lack of animacy

Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL 2016, pages 74–83,
San Diego, California, June 16, 2016.

opposition in pronouns, some specific features of syntactic binding for reflexives in Russian, special cases of "antecedentless" pronouns (when pronouns show semantic reinterpretation) and others. We also focus on some issues that are common for other languages, such as syntactic ambiguity in the case of NP embedding and some cases of referential conflicts. In Section 5, we present our conclusions.

2 Anaphora resolution error analysis: background

Previous studies on anaphora and coreference resolution errors classified system mistakes using different criteria. For instance, Kummerfeld and Klein (2013) consider only deficiencies in the structure of coreferential chains, such as missing/additional mention, span errors, etc. Some studies investigate precision and recall errors in coreference resolution (Uryupina, 2008; Martschat and Strube, 2014) and report particularly difficult cases, namely, resolving 1st and 2nd person pronouns, identifying and linking the names of organizations, and interpreting specific semantic relations, such as meronymy, hyponymy and hyperonymy.

Few works focus specifically on pronominal anaphora resolution mistakes. Barbu (2002) investigates the performance of several anaphora resolution systems and distinguishes errors regarding pronoun types (personal, possessive, reflexives), distance between an anaphor and its antecedent and syntactic function of the referring expressions. Evans (2002) presents a modified system for anaphora resolution in English and defines more detailed error types, such as pre-processing mistakes (syntactic parsing, erroneous encoding or incorrect annotation of training data), non-trivial syntactic cases (number and gender disagreement), distant antecedents, specific types of anaphora (verbal anaphora, cataphora, inferred antecedent, event anaphora) and referential ambiguity. Both studies show that incorrect syntactic processing and distant antecedents have a considerable impact on the accuracy of the system.

Unfortunately, very few studies examine error types in Slavic languages, although in this field we might expect specific mistakes, since Slavic languages pose particular challenges in anaphora resolution due to a rich morphology and free word order (Toldova et al., 2015).

3 Data and Systems: RU-EVAL-2014

3.1 Description of the evaluation campaign

RU-EVAL-2014 was the first evaluation campaign that measured the performance of anaphora and corefence resolvers designed for Russian. It relied on similar evaluation events: MUC-7 (Chinchor and Hirschmann, 1997), EVALITA (Poesio and Uryupina, 2011), ARE (Orasan et al., 2008), SemEval (Recasens et al., 2010), CoNLL-2011/2012 (Pradhan et al., 2011; Pradhan et al., 2012). The aim of the campaign was to assess the state-of-the-art in the field for Russian. The majority of teams dealing with Russian are working with disjoint models (cf. RU-EVAL events on pos-tagging and parsing). This leads to a high diversification of standards and annotation schemes. Thus, evaluation principles of the previous campaigns for other languages had to be adapted for this RU-EVAL event, taking into account specific conditions for developing Russian anaphora resolvers. For the evaluation campaign, the gold standard corpus, the Russian Coreference Corpus (RuCor), was created (Toldova et al., 2016).

3.2 RuCor

RuCor consists of two parts, manually annotated for pronominal anaphora and coreference resolution tasks: the learning set and the evaluation set, 185 texts (200 000 tokens) in total. It is comprised of publicly available texts of different genres (from 5 up to 100 sentences): news (45%), essays (21%), fiction (18%), scientific articles (11%) and blog posts (5%). Each text was manually annotated by two annotators, then the annotation was checked by a supervisor. The corpus also contains automatic morphological annotation. The set of tools, developed by S. Sharoff for Russian was used, which includes a tokenizer, a TreeTagger-based part-of-speech tagger (Schmid, 1994), and a lemmatizer, based on CSTLemma (Jongejan and Dalianis, 2009).

For coreference relations, NPs referring to concrete entities were annotated. Toldova and colleagues (2016) also annotated different types of non-referential expressions that were not taken into consideration in the evaluation procedure, such as pred-

icative and sitive NPs. As for anaphoric relations, four types of pronouns were annotated: 1) 3rd person pronouns (including 3rd person possessive *jego* 'his/its', *eje* 'her', *ih* 'their'), 2) relative pronoun *kotoryj* 'which', 3) reflexive pronoun (*sebya* 'oneself' and a possessive reflexive pronoun *svoj* 'oneself's'), 4) headless demonstratives. The latter were not taken into consideration in anaphora evaluation. Generic and abstract NPs were annotated if they served as antecedents for those pronouns. The annotation provides morphosyntactic characteristics of an NP (full noun group or a pronoun). In NPs containing modifiers, the semantic head of the group is additionally marked, similarly to MUC-7 methodology (Hirschmann, 1997). All the potential heads are annotated. For example, two heads are annotated for an NP *[[professor] [Vagner]]* (person's occupation and surname). There are several possible analyses for this NP: some systems consider only 'professor' to be the NP head, others —the surname. Moreover, some systems link a pronoun to the full NP, while others link it only to the NP head. Thus, the annotation of several potential heads enables to compare systems with different syntactic and coreferential models.

3.3 Participants

Results presented by six different systems were evaluated in the competition. Originally, there were more participants, but as some systems did not manage to analyze the whole evaluation set, they were excluded from further consideration. The final participant lineup was as follows.

- **An@phora**, a system, developed by M. Ionov and A. Kutuzov. The team presented three different runs: one for rule-based approach, one for a Random Forest algorithm and one for a hybrid algorithm (Kutuzov and Ionov, 2014).

- **Compreno**, a linguistic processor, developed by the ABBYY Corporation. It is built upon a self-developed ontology and widely uses semantic analysis. The system provides deep syntactic analysis, using dependency parser designed by this company (Bogdanov et al., 2014).

- A machine-learning based system presented by the Institute of System Analysis, below it is referred to as **ISA** (Kamenskaya et al., 2014). The system developers make use of semantic role labeling to improve its performance.

- A system presented by the Open Corpora project (referred to below as **OC**). It uses Tomita-parser for NP extraction and MaltParser for shallow syntactic parsing (Protopopova et al., 2014).

- **Phenomena**, a machine learning based system, developed individually by S. Ponomarev. It relies heavily on semantic and ontological relations and applies a logistic regression classifier. It involves morphological and syntactic analysis provided by the Tomita parser[1].

- **SemSyn**, a rule-based system, built around the syntactic parser (Boyarski et al., 2013).

3.4 Evaluation

In the pronominal anaphora resolution task, performance on only 3 types of pronominal NPs was evaluated: 1) 3rd person pronouns, 2) the relative pronoun (*kotoryj* 'which') and 3) reflexive pronouns. The zero anaphora was not evaluated. As in Evalita-2011 (Poesio and Uryupina, 2011), we used a weak criterion for antecedent identification. It was not required to link a pronoun to its linear closest non-pronominal antecedent. We treat as true positives the pair of a pronoun and any mention belonging to the same coreference chain which matches the corresponding mention in the gold standard. For instance, in (1), the following pairs: '*him – Vagner*', '*him – professor*' or '*him – he*' are allowed.

(1) *I do not know [Vagner]$_i$ well. Nevertheless, [the [professor]]$_i$ / [he]$_i$ was living nearby, I had met [him]$_i$ just twice.*

The evaluation was based on the principle of lenient matching of NPs: a system antecedent matches an NP in the gold standard corpus (GS) if it includes one of possible heads annotated for this gold standard NP. This makes it possible to compare the results of the systems that differ in principles of antecedent mark-up (cf. NP heads vs. full NPs vs.

[1]Properties of this system are presented in a blog post: https://habrahabr.ru/post/229403/

partial NPs). For example, for the NP *[Professor] [Vagner]* the responses, *professor*, *Vagner* or *professor Vagner* are considered correct. However, the head mismatch in case of embedded NPs as in system response *sumku [mamy]* 'moms bag' for gold standard NP *[mamy]* 'mom' is treated as an error.

We conducted our error analysis based on the systems' responses in the evaluaton set (85 texts, 1600 chains, 2300 pairs). Most of the systems carried out several runs with precision ranging from 36% to 82%. The results are displayed in Table (1).

Run	Algorithm type	P	R	F-measure
sys1	rule-based+onto	0.82	0.70	0.76
sys2	rule-based	0.71	0.58	0.64
sys3	rule-based	0.63	0.50	0.55
sys4	logreg+onto	0.54	0.51	0.53
sys5	svm+sem	0.58	0.42	0.49
sys6	decision tree	0.36	0.15	0.21

Table 1: Evaluation results of RU-EVAL-2014

We present all of the runs. The variation in the results for different runs of one system is not as significant as difference between systems, in spite of the different algorithms employed in different runs.

The rule-based runs generally show better results than those based on machine learning techniques; the top three results are achieved by rule-based systems. Incorporating semantics into analysis leads to better results. The runs involving semantic role labeling, named entity recognition or ontological information achieve higher F-measure scores.

4 Comparative error analysis

The anaphora resolution systems presented in the previous section are a representative sample of the state-of-the-art for anaphora resolution in Russian. Therefore, by analyzing the errors they make, we can uncover remaining challenges in anaphora resolution and analyze qualitative differences between the systems. The results of such an analysis will deepen our understanding of anaphora resolution and will suggest promising directions for further research.

4.1 Error rate analysis for different pronoun classes

In the preliminary analysis, we categorized each error by the pronoun type. Our hypothesis was that syntactic position of a pronoun could influence the error rate. Thus, we distinguished the following classes of anaphoric pronouns: 3rd person pronouns in subject position (nominative case, ***ana_nom***), in direct object position (pronouns in the accusative case, ***ana_acc***), anaphors in prepositional phrases (***ana_pp***[2]) and those in other argument positions (***ana_other***). We also treated possessive 3rd person pronouns (***ana_poss***) as a separate class. As for reflexive pronouns, we split them in two classes: reflexive pronouns proper (***refl***) and reflexive possessives (***refl_poss***), since Russian possessive reflexives have some specific features (cf. Paducheva (1985), see also sections 4.2 and 4.3 for details). Relative pronouns (***rel***) constitute the last class.

Raw frequencies of different pronoun types are presented in table 2. General statistics on error rate is presented in Table 3.

pronoun type	raw frequency
ana_nom	640
ana_acc	217
ana_pp	195
ana_other	174
ana_poss	298
refl	126
refl_poss	294
rel	357
total	2301

Table 2: Statistics on pronoun types

	sys1	sys2	sys3	sys4	sys5	sys6
ana_nom	0.20	0.33	0.43	0.46	0.44	0.72
ana_acc	0.27	0.36	0.43	0.58	0.5	0.75
ana_pp	0.21	0.39	0.46	0.53	0.45	0.77
ana_other	0.23	0.35	0.44	0.45	0.41	0.69
ana_poss	0.20	0.35	0.46	0.42	0.47	0.68
refl	0.20	0.34	0.52	0.83	0.86	0.65
refl_poss	0.17	0.29	0.41	0.55	0.44	0.60
rel	0.19	0.29	0.43	0.55	0.57	0.71
mean	0.21	0.34	0.45	0.55	0.53	0.69

Table 3: Precision error rate for different pronoun types

The raw error rate for different pronoun types depends on system's general performance rather than on the pronoun type. We normalized the error rate

[2]Russian personal pronouns have a special stem starting with *n-* in prepositional context (c.f. *vizhu ego* 'saw him' vs. *pokazal na **n**ego* 'point at him').

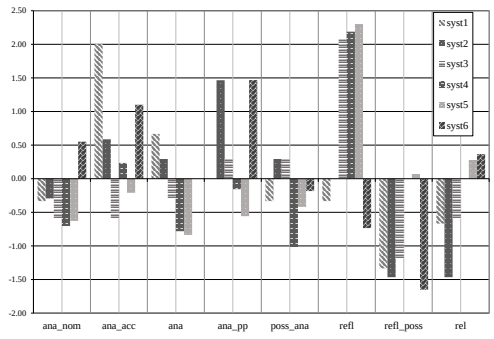

Figure 1: Diagram for different error types across systems

over the system's error rate mean and calculated the deviation (see the comparative diagram in Figure 1). As it can be seen from the diagram, the least problematic cases are possessive reflexives, relatives and 3rd person pronouns in nominative case. The most difficult is the resolution of personal pronouns in accusative case. There is also a tendency for the systems that handle syntactic anaphora (reflexives and the relative pronoun) quite well to have more mistakes in cases of 3rd person pronouns. On the contrary, the systems that are poor at reflexive pronouns analysis outperform the syntax-oriented systems in the discourse anaphora resolution.

We expected that syntactically regulated pronouns (e.g. reflexive and relative pronouns) would be less problematic. This hypothesis was supported, e.g. by (Barbu, 2002), where reflexives are absent in error statistics due to the extremely low rate of such mistakes. But for Russian, contrary to our expectations, the rate of errors in reflexive pronouns had the maximum range across the systems. The high error rate in reflexives is due to the fact that some systems do not take into account the binding theory (see section 4.3.1). However, even those systems that do have a syntactic reflexive resolution model still make mistakes: the lowest error rate for reflexives and possessive reflexives is 20% and 17% for these two types respectively.

It is worth mentioning that relative pronouns in NPs like *kamen, kotoryj* 'the stone which', where the head noun controlling relative clause and the adjacent relative pronoun are coindexed, also present problems for anaphora resolvers (17% error rate for the best result). The high error number in such cases is due to syntactic homonymy in case of embedded NPs as heads, like in *oblomok kamnja, kotoryj* 'a piece of the stone, that', where two anaphoric pairs are possible 'piece – that' or 'stone – that' (see 4.3.2).

The personal pronouns in Russian (as in other languages) are the most difficult issue. The basic criterion for these pronouns is the antecedent – pronoun agreement in morphological features (Jurafsky and Martin, 2009, p. 803). The Russian error analysis reveals issues with this criterion for morphologically rich languages. A lot of systems' mistakes arose due to particular types of morphological homonymy in both pronouns and antecedent forms (cf. *im* –'he.INSTR' vs. 'they.DAT', NOM.PL vs. GEN.SG in feminine nouns), absence of animacy opposition in pronouns etc. (see 4.2 for details).

4.2 Morphological errors

In this section, we will analyze errors that arise due to specific morphological properties of Russian.

In agglutinative languages, morphology usually provides an additional cue for correct anaphora resoloution (see Soraluze et al. (2015) for Basque), but this is not the case for Russian, as it is relatively more flective and tends to express grammatical meanings cumulatively.

Russian personal and possessive pronouns agree with their antecedent in person and number. The third person singular pronouns and possessives also agree in gender (feminine, masculine and neuter in nominative, neuter and masculine are neutralized in oblique cases). In contrast to English, reflexive pronouns *sebja* and *svoj* do not agree in gender and number with the antecedent, and animacy is not marked in any pronouns.

This animacy deficiency together with the neutralization of some gender contrasts in the pronoun system cause additional problems for anaphora resolvers. Animacy deficiency expands the set of potential candidates: systems have to consider both animate and inanimate nouns. For some anaphora resolvers which lack semantic analysis, it is difficult to rule out potential antecedents of the wrong animacy. This was previously reported by Ionov and Kutuzov (2014). Still, in some contexts an inani-

mate reading is more plausible than the animate one. Locative contexts such as *v nem* 'in him' is such an example. But this fact is ignored by systems that are not using deep semantic processing, cf. 2.

(2) *[Nash proekt]$_i$*[3] *otkroet dveri vsem talantlivym ljudjam. My budem predlagat uchastvovat v [nem]$_{is}$ vsem, i [Grigoriju Perelmanu]$_s$ v chastnosti.'*
'[Our project]$_i$ is opened to all talented people. We will offer everyone to participate in [it]$_{is}$, and [Grigori Perelman]$_s$ in particular.'

The only argument of verbs such as *udastsja* 'manage, succeed', on the the contrary, is more likely to be animate, like the pronoun *im* 'them' in 3. Yet, the anaphora resolver links *im* to the intervening *administracii* 'administrations' which is also plural, without considering animacy disagreement.

(3) *. . . [storonniki]$_i$ oppozitsii nachali zahvatyvat [oblastnye administratsii]$_s$ ... [Im]$_{is}$ udalos' . . . '*
'... Opposition [supporters]$_i$ started to occupy [regional administrations]$_s$... [They]$_{is}$ managed...'

Lack of masculine-neuter gender contrasts leads to ambiguity that is difficult to be resolved, cf. the wordform *nego* might be a genitive form of either pronoun *on* 'he' or *ono* 'it'. In the next example (4), the correct interpretation is neuter, but the resolver chooses a more distant masculine antecedent *chelovek* '(a) man'.

(4) *[Chelovek]$_s$, zavedshij [oruzhije]$_i$, dolzhen pozabotitsa o tom, chtoby ot [nego]$_{is}$ ne postradali drugie ljudi.*
'When procuring a [weapon]$_i$, [(a) man]$_s$ must make sure that other people do not fall a victim to [it]$_{is}$.'

Likewise, nominal case-number syncretism misguides the number agreement requirement. An average wordform has 2,5 possible analyses (Toldova et al., 2015), therefore, morphological disambiguation is still problematic. For instance, all feminine nouns and some others have the same wordform for genitive singular and nominative plural, cf. *shkoly* 'school-GEN.SG' or 'school-NOM.PL'. Thus, in (5), due to incorrect morphological analysis, *shkoly* was chosen as the antecedent for *oni* 'they' instead of a more distant plural NP *dva cheloveka* 'two people'.

(5) *[Dva cheloveka]$_i$ upali s kryshi doma. Kazhetsja, [shkoly]$_s$. [Oni]$_i$...*
'[Two people]$_i$ fell from the roof of a building. A [school's]$_s$, it seems. [They]$_{is}$...'

In general, morphological analysis in Russian is done less efficiently than in English. For instance, named entities, such as *Merkel* (Angela Merkel), are often attributed a wrong inflectional class and gender. Besides, even having the correct gender information, some systems choose gender incongruent antecedents for the pronouns, cf. *Vladimirom Putinym* 'Vladimir-INSTR Putin-INSTR' – *ona* 'she'.

To sum it up, Russian rich morphology is an additional source of errors. Some of them are unattested in English anaphora resolution. An anaphora resolver for Russian has to deal jointly with pronoun animacy deficiency, neutralization of gender contrasts in pronouns, nominal case-number syncretism and process novel nouns.

4.3 Syntactic errors

4.3.1 Binding conditions

Errors caused by the violation of syntactic rules were detected in all the systems. The majority of precision mistakes are due to the Binding conditions' violation. Some recall mistakes are due to specificity of binding properties of Russian reflexive pronouns.

Syntactically regulated pronouns, such as reflexives, present no problem for systems in English. E.g. Barbu (2002) reports very low error rate for reflexive pronouns (see also 4.1), as the reflexives do obey the binding conditions: in most cases, the antecedent of a reflexive is in the same clause and occupies the subject position (see (Chomsky, 1981)). In Russian, however, the lowest rate for reflexives is nearly 20% and the range of variation across systems is very high.

Firstly, some mistakes occurred due to difficulties in subject detection. This can be accounted for by free word order and case homonymy in nouns.

[3]Here and further, the index *i* corresponds to the real anaphoric relations, while *s* is the anaphoric links drawn by the system

Secondly, reflexives in Russian can have antecedents in another clause. Russian reflexives (*sebja, svoj*) allow long distance binding, when they occur in infinite clause or within an NP, since PRO and the NP specifier are transparent for binding (Rappaport, 1986). This often prevents the system from finding the correct antecedent and the participants even ignored reflexive pronouns in embedded infinitive clauses. For example, in (6) the system did not find *svoj* in the infinitive clause, although it has a unique antecedent in the same sentence.

(6) *[Ona]$_i$ vyezzhala redko i [∅]$_i$ umela [PRO]$_i$ zastavit' vysoko tsenit' [svoi]$_i$ poseschenija.*
'[She]$_i$ came out rarely and [∅]$_i$ knew, how to [PRO]$_i$ make others appreciate [her]$_i$ visits.'

There are cases of cataphoric usage of the reflexive possessive *svoj* (in 5% of the contexts). In this case systems fail to match this pronoun as in (7), or incorrectly bind it to the antecedent in the preceding sentence as in (8):

(7) *Za [svoju]$_i$ desjatiletnuju istoriju [kompanija]$_i$ sumela stat glavnym poiskovikom.*
'During [its]$_i$ 10 year history, [(the) company]$_i$ managed to become the main search engine.'

(8) *Zapretit' pravjaschuju partiju predlozhili [deputaty]$_s$. V [svoju]$_{is}$ ochered', [mestnyj parlament]$_i$ podkontrolen pravjaschej partii.*
'[(The) deputy]$_s$ suggested to ban the governing party. In [its]$_{is}$ turn, [(the) local parliament]$_i$ is under the control of the governing party.'

Incorrect binding is attested for personal pronouns as well. According to (Chomsky, 1981), personal pronouns are not bound within their local domain, i.e. this pronoun cannot have an antecedent within the same clause.

(9) *[Sasha]$_i$ ljubit [ego]$_{*i}$ / [sebja]$_{i/*j}$*
'[Sasha]$_i$ loves [him]$_{*i}$ / [himself]$_{i/*j}$.'

Applying the strategy of the nearest antecedent that matches the pronoun grammatical features, some systems choose the antecedent in the same clause, although such a decision leads to an ungrammatical interpretation. On the contrary, several participants bind reflexives to a referring group outside their local domain:

(10) *Eto pokazhet nashe otnoshenije k ["ottsu narodov"]$_i$ i tem, [kto]$_s$ pytaetsja [ego]$_{is}$ vykopat'.*
'This will show our attitude to ["the father of nations"]$_i$ and those, [who]$_s$ try to dig [him]$_{is}$ out.'

(11) *On$_i$ ne pozvolil sebe$_{is}$ i legchajshego nameka.*
'He$_i$ did not afford himself$_{is}$ the slightest hint.

Moreover, there are cases of recall mistakes for the reflexive *sebja* in a certain type of idiomatic expressions where it functions not as a proper verb argument, but rather as a middle voice marker, e.g. *pokazat sebja* - 'to come up', *vesti sebja* - 'to behave' (cf. "missing antecedents" type of errors for idiomatic use of pronouns in (Evans, 2002)). Though it is arguable, whether the pronoun has to be linked with the corresponding subject NP, we found out that such a non-standard use of *sebja* caused a number of mistakes for systems.

The high rate of deficiencies in reflexive anaphora resolution highlights the fact that Russian has some specific issues in binding condition modeling. Therefore, the anaphora resolvers need particular heuristics and deeper syntactic analysis in order to handle cases of cataphora and long distance binding.

4.3.2 Parsing errors

Incorrect syntactic parsing influences the results as well. Firstly, we observe errors in matching NP boundaries, especially for NPs with dependent genitive groups, such as *pomoschnik presidenta* 'president's assistant' or *zdanie ministerstva* 'the building of (the) ministry' (the genitive groups are underlined). Several systems incorrectly matched genitives only, ignoring the preceding head of the group and chose them as antecedents.

Secondly, many participants did not recognize multi-word parenthetical words and treat them as PPs or NPs. Consequently, the systems consider the nouns within these expressions as antecedents, in particular, when the nouns appeared at the shortest distance. For example, in (12) the noun *vzgljad* 'view, opinion' is linked to an anaphor, since it is the nearest candidate which agrees in number and gender with the pronoun.

(12) *Na [moj vzgljad]$_s$, [on]$_s$ dolzhen vypolnjat neskolko trebovanij.*

'In [my opinion]$_s$, [he]$_s$ should fulfill some requirements.'

4.3.3 A case of NP embedding

A frequent source of errors is NP embedding. There are two potential antecedents in complex NPs: e.g. possessor vs. full NP in a possessive construction, or NP in a prepositional phrase vs. full NP with a prepositional phrase. In the NP *zdanije ministrestva* 'building of the Ministry' both the possessor and the full NP are potential antecedents for *jego* 'its/his'. The possessor antecedent is a less frequent case, but it is closer to the pronoun. Thus, it is a source for precision mistakes. The same applies to embedding of NPs with prepositional phrases as in *[nash zelenyj sad [nad rekoj]]* '[our green garden [by the river]]'. Especially, it affects the selection of antecedents for relative pronouns(see 4.1). Additionally, grammatical ambiguity influences the correct analysis of such constructions.

4.3.4 Distant antecedent

All the systems limited the position of a potential antecedent to a window of a certain size. If the actual antecedent is located beyond this window, it is ignored by the system. This leads to errors in distant antecedent cases (when antecedents occurred more distant than 2 sentences in the text prior to the pronoun or in the previous paragraph).

According to some reports (Kutuzov and Ionov, 2014; Kamenskaya et al., 2014), setting the maximal window size improves the performance of the system considerably. However, there are rare cases when no appropriate antecedent is located within the fixed window. In (13) reflexive *soboj* should be linked to the personal pronoun *oni*, but instead it is connected to the preposition *pered* 'in front of', which is incorrectly analysed as a homonymous noun *pered* 'front'. Thus, the closest agreement matching antecedent is chosen (a potential antecedent is between a pronoun and its real antecedent), which leads to a precision error.

(13) *[Ljudi]$_i$, nastroennye ekstremistski, [oni]$_i$, kak pravilo, ljudi ogranichennye i ne otdajut sebe otchet v tom, chto dazhe esli, kak [oni]$_i$ dumajut, [oni]$_i$ stavjat [pered]$_s$ [soboj]$_j$ blagorodnye celi, to, sovershaja terroristicheskie [akty]$_s$, [oni]$_i$s otdaljajutsja...*

'[Extremists]$_i$ are, as a rule, very simple-minded and do not realize that even if [they]$_i$, as [they]$_i$ think, have noble ideas [in front of]$_s$ [themselves]$_i$, by committing terroristic acts [they]$_{is}$ move away...'

Thus, there is a tendency for Russian systems to overestimate the linear distance factor for an antecedent, which shows a lack of salience based models for the anaphora resolution task.

4.4 Opaque or pleonastic antecedents

One of the essential issues for the anaphora resolution task is to distinguish the cases of pronouns that have no antecedent (cf. Evans (2002)). For English and some other European languages, expletives present such a problem. As for Russian, there is no obligatory subject in a clause. Impersonal, indefinite-personal, zero pronoun (pro-pronoun) clauses are possible. However, there are special cases of pleonastic antecedents or cases of non-referential pronouns.

Firstly, there are pronouns used in idioms and lexicalized constructions such as in *Vot to-to i ono*_3sg.Neut.PRON 'Here we go', or in honorific terms as in *Jego*_3SG.PRON.POSS *prevoshoditelystvo* 'His excellency'.

Secondly, the pronoun *svoj* has a lexical meaning 'own', so it does not need an antecedent in such cases as in *Svoja*_REFL.POSS *rubashka blizche k telu* 'self before all'.

Standard cases of discontinuous, inferred and implicit antecedents are another source of precision and recall mistakes for Russian. The former are the cases when a plural pronoun refers to two different discourse disjoint NPs and becomes a new group referent (c.f. two arguments of a verb as in 'Peter met John and they...'). The other types of precision errors in case of number disagreement is the so-called associative plural as in *Masha obizhaetsya chto my ih ne zovem* 'Mary takes offence that we don't invite them (Mary and her friends)'. Thus, there are specific cases of pronoun semantic re-interpretation as non-anaphoric elements (as in *svoj* as 'own' or *kotoryj*) and cases of opaque antecedents (e.g. in the associative plural) that affect the anaphora resolution precision.

5 Conclusion

In this study, we have examined different error types that are characteristic for Russian anaphora resolvers. Russian, as a relatively underresourced language with rich morphology, poses challenging issues, such as a lack of animacy distinctions in pronouns, morphological ambiguity, specific binding conditions and particular cases of non-referential pronouns and opaque antecedents. These issues are relevant for all systems which participated in RU-EVAL-2014 evaluation campaign, despite the difference in their approaches and models. Our findings show that language-specific properties require a joint fine-grained analysis of morphology, syntax and semantics, as well as particular rules for some phenomena, such as binding, in order to achieve efficient anaphora resolution for Russian.

Acknowledgments

The reported study was supported by the Russian Foundation of Basic Research, research project No. 15-07-09306 "Evaluation benchmark for information retrieval".

References

C. Barbu. 2002. Error analysis in anaphora resolution. In *LREC*.

A.V. Bogdanov, S.S. Dzhumaev, D.A. Skorinkin, and A.S. Starostin. 2014. Anaphora analysis based on {ABBYY} compreno linguistic technologies. 13(20).

K. K. Boyarski, E. A. Kanevskij, and Stepukova A. V. 2013. Vyjavlenie anaforicheskih otnoshenij pri avtomaticheskom analize teksta [in russian, 'detection of anaphoric relations in automatic text processing']. *Nauchno-tehnicheskij vestnik informatsionnyh tehnologij, mehaniki i optiki*, 5(87):108–112.

N. Chinchor and L. Hirschmann. 1997. Muc-7 coreference task definition, version 3.0. In *Proceedings of MUC*, volume 7.

N. Chomsky. 1981. *Lectures on Government and Binding*. Foris.

R. Evans. 2002. Refined salience weighting and error analysis in anaphora resolution. *Proceedings of Reference Resolution for Natural Language Procesing*, pages 51–59.

L. Hirschmann. 1997. Muc-7 coreference task definition. version 3.0. In *Proceedings of the 7th Message Understanding Conference (1997)*.

B. Jongejan and H. Dalianis. 2009. 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*, pages 145–153. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

M.A. Kamenskaya, I.V. Khramoin, and I.V. Smirnov. 2014. Data-driven methods for anaphora resolution of russian texts. 13(20).

J. K Kummerfeld and D. Klein. 2013. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 265–277.

A. Kutuzov and M. Ionov. 2014. The impact of morphology processing quality on automated anaphora resolution for russian. In *Computational Linguistics and Intellectual Technologies*, volume 13.

S. Martschat and M. Strube. 2014. Recall error analysis for coreference resolution. In *EMNLP*, pages 2070–2081.

C. Orasan, D. Cristea, R. Mitkov, and A. H. Branco. 2008. Anaphora resolution exercise: an overview. In *LREC*.

E. V. Paducheva. 1985. *Vyskazyvanie i ego sootnesnnost' s dejstvitel'nostju [In Russian, 'Utterance and its interrelationship with reality']*. Ripol Klassik.

M. Poesio and O. Uryupina. 2011. Anaphora resolution task at Evalita 2011. In *Working Notes of EVALITA 2011*.

S. Pradhan, L. Ramshaw, M. Marcus, M. Palmer, R. Weischedel, and N. Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 1–27, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, and Y. Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 1–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

E.V. Protopopova, A.A. Bodrova, S. A. Volskaya, I. V. Krylova, A. S. Chuchunkov, S. V. Alexeeva, V. V. Bocharov, and D. V. Granovsky. 2014. Anaphoric annotation and corpusbased anaphora resolution: an experiment. 13(20).

G. C Rappaport. 1986. On anaphor binding in russian. *Natural Language & Linguistic Theory*, 4(1):97–120.

M. Recasens, L. Màrquez, E. Sapena, A. Martí, Ma. Taulé, V. Hoste, M. Poesio, and Y. Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

H. Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer.

Ander Soraluze, Olatz Arregi, Xabier Arregi, and Arantza Díaz de Ilarraza. 2015. Coreference resolution for morphologically rich languages. adaptation of the stanford system to basque. *Procesamiento del Lenguaje Natural*, 55:23–30.

S. Toldova, A. Roytberg, A. Ladygina, M. Vasilyeva, I. Azerkovich, M. Kurzukov, G. Sim, D. Gorshkov, A. Ivanova, A. Nedoluzhko, et al. 2014. Ru-eval-2014: Evaluating anaphora and coreference resolution for russian. In *Computational Linguistics and Intellectual Technologies*, volume 13.

S. Toldova, O. Lyashevskaya, A. Bonch-Osmolovskaya, and M. Ionov. 2015. Evaluation for morphologically rich language: Russian nlp. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, page 300. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

S. Toldova, Yu. Grishina, A. Ladygina, M. Vasilyeva, G. Sim, and I. Azerkovich. 2016. Russian coreference corpus. In Francisco Alonso Almeida, Ivalla Ortega Barrera, Elena Quintana Toledo, and Margarita E. Sánchez Cuervo, editors, *Input a Word, Analyze the World*. Cambridge Scholars Publishing.

Olga Uryupina. 2008. Error analysis for learning-based coreference resolution. In *LREC*.