Features for discourse-new referent detection in Russian

Svetlana Toldova[1] and Max Ionov[2]

[1] National Research University “Higher School of Economics”,
toldova@yandex.ru
[2] Goethe University Frankfurt / Moscow State University,
max.ionov@gmail.com

Abstract. This paper concerns discourse-new mention detection in Russian. This might be helpful for different NLP applications such as coreference resolution, protagonist identification, summarization and different tasks of information extraction to detect the mention of an entity newly introduced into discourse. In our work, we are dealing with the Russian where there is no grammatical devices, like articles in English, for the overt marking a newly introduced referent. Our aim is to check the impact of various features on this task. The focus is on specific devices for introducing a new discourse prominent referent in Russian specified in theoretical studies. We conduct a pilot study of features impact and provide a series of experiments on detecting the first mention of a referent in a non-singleton coreference chain, drawing on linguistic insights about how a prominent entity introduced into discourse is affected by structural, morphological and lexical features.

Keywords: coreference resolution, discourse-new referent, discourse processing, natural language processing, machine learning

1 Introduction

The task of tracking all mentions referring to the same entity in discourse, or coreference resolution task, is essential for many text-mining applications. This task has received much attention over the last years ([21], [15]) and nowadays it has achieved a rather high precision. Although current machine learning approaches to coreference resolution perform quite well, many studies have sought to improve the quality, not only via improving the basic algorithms but also via taking into consideration various theoretical assumptions from the discourse models of referential choice (cf. the hierarchy of referential accessibility, e.g. [1]). Among others, there are issues such as discourse-new recognition. The decision on whether an NP is introducing a new entity into discourse could improve the quality of coreference resolution ([10], [20]). Moreover, a particular type of an introductory NP could be a clue to the discourse role of a referent as to whether it is an entity that is the main topic of a long discourse span or it is an occasional one.

There is a comprehensive analysis of different features for discourse new (DN) detection for English ([16], [28], [19], [10], etc.). However, there is much less investigation concerning the first mention detection for languages without articles such as Russian. While for the former the only task is to decide whether an NP introduces a new referent in spite of an overt definite marker, for the latter the task is to differentiate among nearly all the NPs (except anaphoric pronouns and NPs with demonstratives) as to whether they introduce a new referent or not. This task is complicated for less resourced languages as there is no freely available high quality syntactic parsers or rich semantic resources such as WordNet. Thus, one could hardly rely upon some complicated syntactic NP properties or semantic NP heads relatedness. However, there is some theoretical research (cf. [2], [26]) concerning the specific NP structure features and lexical features that serve as markers for introducing a new referent. Moreover, there are special markers for introducing prominent referents (referents in focus of attention, cf. [6]) into discourse. The question is whether these features could be helpful within a shallow machine learning approach for discourse-new mention detection.

In our work we conduct a pilot study of discourse-new detection in Russian. We take a subset of basic features for discourse-new mention detection for English and enrich it with features mentioned in the theoretical literature (mainly for Russian). The task is to check whether the theoretically assumed features really work. We suggest that the analysis of these features impact on the first mention vs. repeated mention classification given that the singletons (NPs referring to the entities mentioned only once) are filtered out in the previous stages of analysis.

The paper is structured as follows. In section 2 we present the prerequisites for our experiment. In section 2.1 the overview of general approaches to coreference resolution task is given. In section 2.2 the approaches for the first mention detection in English are discussed. The subsection 2.3 deals with various theoretical investigations of first-mention NP properties. Section 3 is devoted to the investigation of discourse-new detection features, their distribution in corpus and their correlation with the introductory NPs. In section 4 we describe our experiments on DN mention detection and suggest the analysis of the features impact.

2 Background

2.1 Coreference resolution task and discourse new referent detection. Task settings

Coreference resolution is the task of grouping NPs referring to the same entity (mentions or referring expressions) into disjoint sets. An example is given in example 1, where mentions from three sets are marked with brackets and corresponding indexes:

(1) . . . Probovali sravnivat' text [zapisnoj knižki]$_{a1}$ s rukopisjami [Nahimova]$_{a2}$.

<... >issledovatelej zainteresovala [ešjo odna [jego]$_{a2}$ zapisnaja knižka]$_{a3}$.

(they) tried to compare (the) text of (the) [notebook]$_{a1}$ with (the) [Nahimov]$_{a2}$'s manuscripts. The researchers were interested in [another ([his]$_{a2}$) notebook]$_{a3}$.

This example includes different types of referential expressions: the named entity (*Nahimov*), the possessive pronoun *jego* referring to this named entity, and NPs referring to two different entities of the same ontological class ('notebooks'). The NP$_{a1}$ *zapisnaja knizhka* (notebook) has no overt markers as to whether it is the first mention of an entity, a repeated one or if it is a generic use of 'notebook' referring to a class of entities. Such ambiguity is a specific challenge for coreference resolution in languages without articles. Another difficulty is that NP$_{a1}$ and NP$_{a3}$ have a common part *zapisnaja knizhka* (notebook), though they refer to different entities. However, the latter has *esche odna* (one more) as a specific introductory modifier. Thus, there are cases where the DN detection might be helpful. Moreover, there are special lexical DN markers that serve as a signal for introducing the reference in discourse (cf. [17], see section 2.3 and section 3.4 for details).

The procedure for establishing coreference in the text such as that above can involve different types of information ranging from simple formal features such as token distance, NPs and NP heads equivalence, morphological congruence, syntactic features etc. ([16], [15] etc.) up to high level discourse and semantic features ([13], [19], [21] [10]) and others). Moreover, it can include the task of separating singletons (NPs for entities mentioned only once in discourse) from coreferential NPs ([24], [28]).

Further knowledge that could influence the coreference resolution performance is discourse-new mention detection (see discussions in [16], [19], [10] etc.). This paper focuses on this task.

2.2 Overview of discourse-new detection algorithms for English

The majority of previous works on discourse-new (DN) detection deal with English. One of the most detailed overview of various approaches to this task is given in [20]. The approaches suggested in [3], [19], and [28] are among them. Though Ng and Cardie ([16]) reported that the incorporation of DN recognition into coreference resolution systems did not affect their general performance, further testing of the features from [16] as well as the development of other algorithms shows that the way in which DN detection is combined with the basic coreference resolution module also matters (see [20], [10] for details). In this paper we put the question of the DN utility aside and focus on the features that are used for DN detection.

The motivation for developing the algorithms for the DN detection task in English was the corpus study carried out by Viera and Poesio ([22]) where they reported that 52% of definite descriptions are discourse new. Thus, the main issue of DN recognition in English (apart from detecting indefinite descriptions) is to identify the class of definite descriptions. Vieira and Poesio in [29] suggested an algorithm for identifying five categories of definite descriptions that are licensed to occur as the first mention, on semantic or pragmatic grounds: (i) semantically functional descriptions ([14]) such as *the best* or *the first*; (ii) "descriptions serving as disguised proper names, such as The Federal Communications Commission"; (iii) predicative descriptions, including appositives and NPs in certain copular constructions, e.g. *Mr. Smith*, *the president of* . . . , etc.; (iv) "descriptions established (i.e., turned into functions in context) by restrictive modification ([14]) as in *The hotel where we stayed last night was pretty good*; (v) larger situation definite descriptions ([8]), i.e., definite descriptions like *the sun*, *the pope*, etc., which denote uniquely on the grounds of shared knowledge about a situation (Löbner's 'situational functions')".

Regarding the various features used for DN mention detection for English, a comprehensive set of features was analyzed and tested in [16]. They used 37 features to identify anaphoric and not-anaphoric NPs taking into consideration the information proposed in the previous works. Their set of features is based on grammatical features of an NP (whether an NP is a certain type of pronoun, a proper noun etc. or whether an NP has a modifier of a certain grammatical type), some string relation features (cf. 'lexical features') based on the identity of an NP or of its head to a preceding NP or the head of a preceding NP. Another class of features concerns some special constructions such as an NP with a relative clause, an NP with the superlative modifier, etc. These constructions require a definite marker in an NP, regardless of whether the NP was used anaphorically or not. This class of features is not relevant for Russian (see section 2.3). Ng and Cardie also took into consideration semantic relatedness and NP sentence position. Another type of "constructional" features mentioned in the literature refers to the question of whether an NP is in an appositive or predicative construction ([3]). However, these features need special syntactic analysis for Russian, which is beyond the scope of the present study.

The core aim of our experiment is to test the theoretical assumptions related to the referent first mention NP in Russian and find out whether they can be helpful for the task. Hence, even a subset of formal features is enough to set the baseline for this task. We took into consideration only string relations and grammatical features as the basis for our experiment (see section 3.2 for details).

2.3 Discourse-new referent recognition in Russian: theoretical accounts

As has been shown in section 2.1, the DN detection task for Russian (as a language without articles) differs significantly from that for English. In Russian, there are no overt grammatical markers of definiteness. Almost all the NPs (excluding anaphoric pronouns, refelexives and NPs with demonstratives) can refer to a newly introduced entity. Moreover, common NPs can have up to three interpretations (e.g. definite, indefinite or generic description, see 1 in 2.1).

One of the approaches for resolving ambiguous interpretations of an NP in article-less languages is to take into consideration the discourse status of the NP itself. This approach goes along with different cognitive-based coreference models and various typological findings. It has deep theoretical motivation, e.g. the hierarchy of referential accessibility suggested in Ariel ([1]) (see also [5], [23] etc.). This hierarchy regulates the choice of anaphoric expressions: the more accessible a referent is, the less informative expressions can be used to refer to it; meanwhile the less accessible entities (e.g. newly-introduced into discourse) need more informative anaphoric expressions. The following hierarchy of NP types: zero < anaphoric pronouns < demonstratives < full NP corresponds to the referent accessibility hierarchy. The notion of accessibility varies through different discourse models, cf. 'topicality' as in [5], salience, activation ([12], [11]), or whether an entity is in the focus of attention or not ([6]).

In [26], the theoretical account for the distribution of different full NP types in discourse is suggested. It is based on the notion of focus of attention (cf. [6]). The licensing of certain NP types for a referent depends on the basis of its discourse properties on whether it is in focus in a particular discourse unit or not.

Arutyunova ([2]) analyses the full NP structure with respect to the referent first-mention / non-first mention description. The main properties of the first-mention NPs specified by Arutyunova are as follows: (i) the NP length (they tend to be longer than the non-first ones); (ii) the number of adjectives (if it is higher than average); (iii) the semantics of adjectives (there is a tendency to use evaluative and qualitative adjectives). She also mentions special predicate types for referent introduction such as existential predications (c.f. features for discourse new descriptions detection suggested by Ng and Cardie ([16])).

These observations are summed up in [26] and [4]. The latter presents a corpus analysis of introductory NPs in mass media texts of a special kind. The relevant features that we have taken into account from this work are special lexical clues for the introduction of a new referent in the focus of attention such as some types of indefinite pronouns, non-identity markers, novelty markers etc.

3 Features analysis

Below we suggest an analysis of the features proposed in the theoretical literature for Russian (some of them coincide with the features discussed above). We rely only on a subset of structural features (see section 2.1 for the discussion) as the basis for the DN mention classifier and explore the impact of the features mentioned in section 2.3.

3.1 Data

Our experiments were conducted on RuCor, a Russian coreference corpus released as a part of the RU-EVAL campaign ([27])[3].

The corpus consists of short texts or fragments of texts in a variety of genres: news, scientific articles, blog posts and fiction. The whole corpus contains about

[3] The corpus may be downloaded on http://rucoref.maimbava.net.

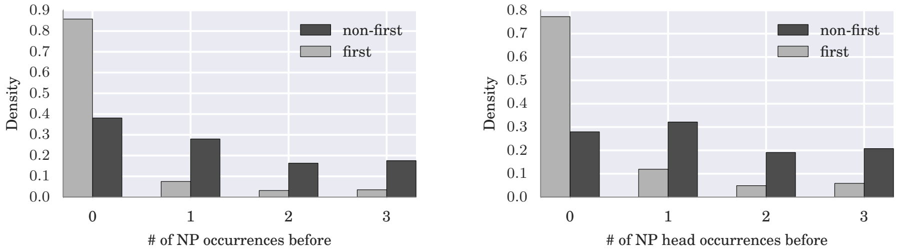
180 texts and 3 638 coreferential chains with 16 557 noun phrases in total. Each text in the corpus is tokenized, split into sentences and morphologically tagged using tools developed by Serge Sharoff ([25]). Noun phrases were obtained using a simple rule-based chunker ([9]). The corpus was randomly split into train and test sets (70% and 30% respectively).

Since anaphoric, reflexive, and relative pronouns cannot be used as first (non-anaphoric) mentions, it seemed more fair to ignore all the instances of these pronouns.

3.2 Basic features

Firstly, we use string relation features. On the one hand, if an NP matches one of the preceding NPs or its head, it is highly probable that the former is a non-first mention. The number of occurrences of an NP or its head in the preceding text can be used as a feature (for English these features were used, for example, in [16]). Using these undoubtedly important features, we are setting the base level for the first mention detection.

Figure 1 shows a distribution of these features in the training set. In most cases there is a low number of preceding NP or head occurrences for the first mention NPs (c.f. more than 80% preceding NP match for non-first mentions).



(a) Distribution of occurrences of the full NP (b) Distribution of occurrences of NP head

Fig. 1. Distribution of string feature values

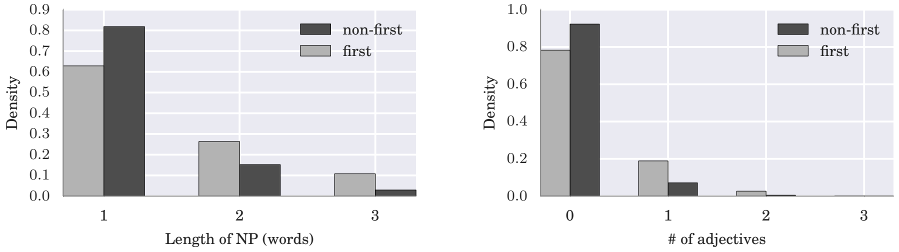
We also used some grammatical features of an NP as basic features (e.g. if the NP is a proper noun, has a demonstrative as a modifier etc.). For the full list see table 4.

3.3 Structural features of first mention NPs

One of the introducing strategies is an extensive description of a new referent within the first-mention NP ([2,4]). Consider example 2 where the new Pope is introduced into a discourse. The introductory NP is composed of five tokens in length and has two adjectival premodifiers.

(2) *76-letnij novyj glava milliarda katolikov.*
The 76 year-old new head of a billion of Catholics.

This assumption entails two classes of useful features: **NP length** (introductory NPs tend to be longer than non-first mentions) and **NP POS structure** (introductory NPs have more adjectival modifiers). Distributions for those features are presented on figure 2. Though the one-token NPs are quite frequent in both classes, the longer NPs are predominately DN referent mentions.



(a) Distribution of NP lengths

(b) Distribution of number of adjectives in NP

Fig. 2. Distribution of structural features

The number of premodified adjectives is higher for introductory NPs relative to the average number of premodified adjectives.

3.4 Special lexical clues

Special attention for two more classes of features which are less discussed in the literature, is deserved. Usually the focus is on the special determiners that mark the identity of NP referents. However, there is a special class of so-called alterators (term used by Palek, [17]). They mark the inequity of an NP referent to any preceding NP referent. Several classes of such alterators were suggested in [26]:

(a) indefinite markers, e.g. *odin* ‘one’, *nekij* ‘a person’;
(b) the inequity markers such as *drugoy*, *inoj* ‘other’ etc.;
(c) the similarity markers such as *takoj* ‘such, *this* kind of’, *podobnyj* ‘analogous’, *pohozchij* ‘similar’ etc.;
(d) markers for elements from a set *odin iz* ‘one of the’;
(e) *ostal’nue* ‘the rest’;
(f) the order of introduction: *pervyj is (nih)* ‘the first’, *vtoroj* ‘the second’, *poslednij* ‘the last’.

Although the alterators do not occur very frequently in discourse, they can be reliable features for discourse-new descriptions detection.

Though Russian lacks articles there are still certain types of indefinite pronouns (specific indefinites, such as *kakoj-to* 'some of', *nekij* 'one of') that also occur only in first-mention NPs (cf. the feature 'indefinite pronouns' in [16]). The same is true for some quantifiers such as *mnogije* 'many of' or *bol'shinstvo* 'the majority' and others (cf. the feature 'quantifiers' in [16]).

There are also certain semantic classes of adjectives (see [4]), such as adjectives denoting the novelty of an entity *novyj* 'new', *nedavno otkrytyj* 'newly discovered' or evaluative adjectives such as 'mysterious', 'strange', 'curious', 'nice', 'modern', 'promising' and others. The corpus analysis provided in [4] has shown that these adjectives are typical for introductory contexts. The last class worth mentioning is a class of so-called classifiers whose nouns have very general semantics such as *chelovek* 'a man', *predmet* 'a thing', etc.

We take the lexical classes enumerated above as features for testing and compile the corresponding lists manually on the basis of lists suggested in the literature.

3.5 Automatically extracted lexical features

The lexemes described above are not very frequently occurring features. In order to enlarge the list of adjectives that are specific for introductory NPs we performed an experiment that aimed to extract them automatically. Since this experiment in relies heavily on our main experiments, its design and results are presented hereafter in section 4.3.

4 Experiments

4.1 Task

To check if the features proposed in the previous sections are adequate for distinguishing discourse-new mentions from recurring ones, we performed a series of experiments. For each experiment we built a classifier with a subset of features.

For our experiments we assumed that each noun phrase in a text belongs to one of the following three classes:

1. First mention (DN): the first mention of the coreference chain,
2. Recurring mention (Non-DN): the mention of an already introduced coreference chain,
3. Singleton: the mention of a referent that appears only once in the text.

Classifying mentions into these classes can be done in a number of ways. The most straightforward model for this task is multiclass classification with mentions as instances and 3 target classes. However, the distinction between the first two classes is different from the distinction between any of these classes and a singleton. In the first case there are two mentions (elements of some coreferential chain) with different discourse roles; however singletons do not usually count as mentions. Therefore, the task of identifying first mentions among all non-singleton mentions and identifying singleton mentions should be treated as two different tasks.

Since the focus of this paper is on the differences between discourse-new and referring coreferent mentions, in our experiments, singleton mentions were filtered out beforehand. Identifying singleton NPs for Russian is a matter of further research. Some approaches of filtering singletons in English are discussed in [28,24].

4.2 Data

As has been mentioned in section 3.1, we use RuCor for our experiment, taking 70% of it as a training corpus and 30% as a testing corpus. Since the dataset is unbalanced (instances of classes are in the ratio 1:4), we performed a balancing operation on a training set. Because of the relatively small size of the corpus we preferred oversampling to undersampling techniques. We chose Borderline-SMOTE1 method ([7]) over several popular oversampling algorithms since on average it yielded best results. Table 1 shows the number of instances before and after balancing.

	No balancing	With balancing
first	3 411	12 738
non-first	12 367	9 868

Table 1. Number of instances before and after balancing (SMOTE ratio=3.626)

Since there is no baseline for the DN-detection task for Russian, we created a simple heuristic baseline classifier with a set of heuristic rules. It works as follows:

- NP is a **first occurrence** if there are no other occurrences of this NP before
- NP is a **recurring occurrence** otherwise

4.3 Features

In our experiments we used several groups of features:

1. String relation features (class “string”)
2. NP structural features (“struct”)
3. Theoretically motivated lexical features (“lists”)
4. Automatically extracted lexical features (“lists”)

Below we describe these features.

String relation features This group contains features that take into account the presence of the NP or its part in the preceding text. As we have seen in section 3.2, most of the instances have low values of this feature. Given this, we may reformulate those features as the following binary features:

- **str_match_before**, number of occurrences of this NP in the text before: 0, less than 2, less than 3, 2 or more;
- **head_match_before**, number of occurrences of the head of this NP in the text before: 0, less than 2, less than 3, 2 or more;
- **is_proper_noun**, true if the whole NP is a proper noun
- **latin**, true if NP contains Latin letters;
- **uppercase**, true if NP is uppercase.

NP structural features This group consists of two sets of features: those concerning the length on NP in tokens and those concerning the number of adjectives in the NP.

More specifically, we have extracted these binary features:

- **len_np**, the length of the NP in tokens: less than two words, more than two words;
- **n_adj**, the number of adjectives in the NP: 0, more than one, more than two;
- **conj**, true if there is a conjunction in the NP.

Pronominal modifiers and theoretically motivated lexical features We manually compiled several lists of words corresponding to theoretical expectations of discourse-new markers[4]. More precisely, we used the following lists:

- Demonstrative pronouns;
- Possessive pronouns;
- General class names: nouns that define a class (*building*, *manager*, etc.);
- Indefinite pronouns;
- New referent introductory adjectives (*contemporary*, *latest*, etc.);
- Non-identity and similarity markers: *another*, *similar*, etc.;
- Common knowledge markers (*famous*, *legendary*, etc.);
- Adjective markers of a discourse role in an NP (*main*, *small*, etc.);
- Subjective markers (*good*, *prestigious*, etc.);

Automatically extracted lexical features To extract the most important adjectives for the classification we performed a univariate feature selection operation using the χ^2 metric and 'bag-of-adjectives' as features: each feature meant the presence / absence of a unique adjective that we encountered in the training corpus. After this procedure we manually cleaned this list by removing the pronouns and words erroneously tagged as adjectives. From the cleaned list we extracted the 50 most important adjectives.

Top 10 adjectives from the list are given in table 2.

[4] Pronouns are a closed grammatical class therefore it may be treated as a list.

#	Adjective	Translation
1	novij	new
2	radioakivnij	radioactive
3	russkij	Russian
4	pervij	first
5	sotsial'nij	social
6	mestnij	local
7	sobstvennij	own
8	global'nij	global
9	nebol'shoj	not-big
10	regional'nij	regional

Table 2. Top 10 adjectives most valuable for classification

4.4 Results

To implement the classifier, we used a Random Forest classifier from the scikit-learn Python library ([18]). Since the test portion of our data set is unbalanced, overall classifier quality is not as important as the quality for the minority class. Results for this class are shown in table 3. For each set experiment we state precision, recall and F1-measure.

	P	R	F1
Baseline	0.505	0.824	0.627
String	0.539	0.770	0.634
String + Struct	0.563	0.698	0.623
String + Struct + Lists	0.573	0.705	0.632

Table 3. Classification results

In table 3 we compare the performance of the classifier using structural features such as NP length and number of adjectives against the baseline and the classifier trained only on the basic feature set. The results show the increase in precision by more than two percent. However, the recall has decreased.

According to the table, the precision for all the features including the lexical lists increases by one percent. The recall has also increased slightly as compared to the classification without lists. Detailed analysis of the results shows the positive influence of those features on the performance. A minimal effect of structural and lexical features can be accounted for by some peculiarities of discourse structure (see 4.6) and the high sparseness of the list features. However, we find it a promising direction for further investigation of specific lexical features.

4.5 Analysis of feature contribution

To measure the importance of each feature we created a logistic regression classifier and trained it on our training data. Each coefficient of a trained classifier showed the impact of the corresponding feature. The results are shown in table 4.

Feature	Class	Coefficient
$str\_match\_before = 0$	string	1.0904
$str\_match\_before < 2$	string	−0.0704
$str\_match\_before < 3$	string	0.1171
$str\_match\_before > 2$	string	−0.4735
$head\_match\_before = 0$	string	1.2394
$head\_match\_before < 2$	string	0.3106
$head\_match\_before < 3$	string	−0.2663
$head\_match\_before > 2$	string	−0.0901
uppercase	string	−0.6143
latin	string	−0.4735
is_proper_noun	string	−0.9239
conj	struct	−0.1072
$len\_np < 2$	struct	−0.2964
$len\_np > 2$	struct	0.3172
$n\_adj = 0$	struct	−0.0161
$n\_adj > 1$	struct	0.0012
$n\_adj > 2$	struct	−0.5231
in_list_refer_to_CommKnowl	lists	0.4179
in_list_adj-top50	lists	0.9041
in_list_role_assess	lists	−0.4250
in_list_NewRef	lists	−0.5883
in_list_non-identity_sim	lists	−0.4523
in_list_possessives	lists	−2.3525
in_list_subjectivity	lists	−1.0180
in_list_class	lists	0.9906
in_list_demonstratives	lists	−1.6133
in_list_indef	lists	1.2672

Table 4: Feature importances for the DN detection task

Our study shows shows that all the classes of features discussed above do matter for the DN mention detection task. For example, an NP length of more than one token has a positive correlation with DN class. The lexical features are not homogeneous. The non-identity and similarity alterators show a negative effect. However, a study of particular examples of this class has shown that this feature needs more precise analysis. The common-knowledge adjectives as well as indefinite pronouns and classifiers are more reliable features than subjectivity adjectives. The automatically extracted lexical features (see section 3.5) work better than manually built lists of subjectivity adjectives.

4.6 Discussion

Some of the most frequent types of precision mistakes are predicative NPs and appositive NPs that we did not take into consideration in our analysis. The fact is that these NP types tend to be long in tokens and include the vast description of an entity as in example 3:

(3) *On yavlyaetsya glavnym sponsorom kluba.*

The other class of cases when the longest NP for a referent is not the first one is a special discourse strategy of a referent introduction in fiction. The first mention is just a proper name without any details (cf. *Mashka* 'Mary') and the second mention is a detailed description of an entity (*Khuden'kaya bol'sheglazaya devochka* '(the) slim big-eyed girl').

The feature **in_list_non-identity_sim** has a negative coefficient. More detailed analysis of the examples shown that these are primarily NPs with the modifiers *takoj* 'such as', *podobnyj* 'similar to'. These cases are highly problematic for annotation for the majority of negative examples that concern abstract NPs with no referent. These are the cases of near-identity. Thus, these modifiers work well with concrete NPs, however, they occur more frequently with non-referential NPs.

The other type of mistakes (recall mistakes) comes from underestimation of one-token NPs. However, this type of DN NP is frequent for two-element chains, while the referents that are mentioned more than twice in discourse tend to be introduced with longer NPs. Though this hypothesis needs further investigation.

Thus, testing theoretically-accounted features and analysis of the cases contradicting with theory reveal new theoretically interesting phenomena and suggest new features for analysis.

5 Conclusions

In the present work we discussed various features used for the first mention detection classification task as a subtask of coreference resolution systems. We presented preliminary research on first mention detection with special emphasis on the Russian language.

The focus was on the analysis of the features used for discourse-new descriptions detection. We analyzed theoretically predicted features (based on typological and cognitive approaches to DN detection) and estimated their contribution. A set of additional features for Russian DN descriptions detection was suggested.

We tested classifiers that can distinguish between first mentions and recurring mentions. We also set a baseline for further experiments and tested special lexical features for referent novelty and inequality marking.

The analysis of the results of this first experiment for DN detection in Russian has shown that the lexical features are quite promising for this task and need further investigation and enhancing. The other promising direction of the research is the correlation of different features with the referent prominence (or the length of the coreference chain). In our future work we are planning to examine the contribution of more elaborated features to this task.

Acknowledgments

The authors would like to thank anonymous reviewers for their helpful comments, the Lomonosov Moscow University students who participated in the corpus markup, and Dmitrij Gorshkov for software support.

This research was supported by grant from Russian Foundation for Basic Research Fund (15-07-09306).

References

1. Ariel, M.: Accessing Noun-Phrase Antecedents. Routledge (1990)
2. Arutyunova, N.: Nomination, reference, meaning. [nominaciya, referenciya, znacheniye] (in Russian). In: Nomination: General Questions. [Nominaciya: obshie voprosi]. Nauka (1980)
3. Bean, D.L., Riloff, E.: Corpus-based identification of non-anaphoric noun phrases. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 373–380. ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA (1999)
4. Bonch-Osmolovskaya, A., Toldova, S., Klintsov, V.: Introductory noun phrases: a case of mass media texts. [strategii introduktivnoj nominacii v teksrah smi] (in Russian) (2012)
5. Givón, T. (ed.): Topic Continuity in Discourse: A Quantitative Cross-Language Study. John Benjamins., Amsterdam (1983)
6. Grosz, B.J., Weinstein, S., Joshi, A.K.: Centering: A framework for modeling the local coherence of discourse. Comput. Linguist. 21(2), 203–225 (Jun 1995)
7. Han, H., Wang, W.Y., Mao, B.H.: Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I, chap. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, pp. 878–887. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
8. Hawkins, J.A.: Definiteness and indefniteness: a study in reference and grammaticality prediction. London: Croom Helm (1978)
9. Ionov, M., Kutuzov, A.: Influence of morphology processing quality on automated anaphora resolution for russian. In: Proceedings of the international conference Dialogue-2014. RGGU (2014)
10. Kabadjov, M.A.: A comprehensive evaluation of anaphora resolution and discourse-new classification. Ph.D. thesis, Citeseer (2007)
11. Kibrik, A., Linnik, A., G., D., Khudyakova, M.: Optimizacija modeli referencial'nogo vybora, osnovannoj na mashinnom obuchenii [optimization of a model of referential choice, based on machine learning]. In: Computational Linguistics and Intellectual Technologies. vol. 11, pp. 237–246. Moscow, RGGU (2012)

12. Kibrik, A.A.: Reference in discourse. Oxford University Press (2011)
13. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Comput. Linguist. 20(4), 535–561 (Dec 1994)
14. Löbner, S.: Definites. Journal of semantics 4(4), 279–326 (1985)
15. Mitkov, R.: Anaphora resolution: the state of the art (1999)
16. Ng, V., Cardie, C.: Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1. pp. 1–7. Association for Computational Linguistics (2002)
17. Palek, B.: Cross-Reference a Study from Hyper-syntax. Universita Karlova (1968)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
19. Poesio, M., Kabadjov, M.A.: A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In: In Proceeding of LREC. pp. 663–666 (2004)
20. Poesio, M., Kabadjov, M.A., Vieira, R., Goulart, R., Uryupina, O.: Does discourse-new detection help definite description resolution. In: In Proceedings of the Sixth International Workshop on Computational Semantics, Tillburg (2005)
21. Poesio, M., Ponzetto, S.P., Versley, Y.: Computational models of anaphora resolution: A survey (2010)
22. Poesio, M., Vieira, R.: A corpus-based investigation of definite description use. Comput. Linguist. 24(2), 183–216 (Jun 1998)
23. Prince, E.F.: The zpg letter: Subjects, definiteness, and information-status. Discourse description: diverse analyses of a fund raising text pp. 295–325 (1992)
24. Recasens, M., de Marneffe, M.C., Potts, C.: The life and death of discourse entities: Identifying singleton mentions. In: Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 627–633. Association for Computational Linguistics, Stroudsburg, PA (June 2013)
25. Sharoff, S., Nivre, J.: The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In: Proc. Dialogue, Russian International Conference on Computational Linguistics. Bekasovo (2011)
26. Toldova, S.: Struktura diskursa i mehanizm fokusirovaniya kak vazhnie faktori vibora nominatsii ob'ekta v tekste (Discourse structure and the focusing mechanism as important factors of referential choice in text) (1994)
27. Toldova, S., Rojtberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., Ivanova, A., Nedoluzhko, A., Grishina, J.: RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. Computational Linguistics and Intellectual Technologies 13 (20), 681–694 (2014)
28. Uryupina, O.: High-precision identification of discourse new and unique noun phrases. In: ACL Student Workshop. Sapporo (2003)
29. Vieira, R., Poesio, M.: An empirically based system for processing definite descriptions. Comput. Linguist. 26(4), 539–593 (Dec 2000)