ISSN 2007-9737

Mention Detection for Improving Coreference Resolution in Russian Texts: A Machine Learning Approach

Svetlana Toldova[1], Max Ionov[2]

[1] National Research Institute “Higher School of Economics”, Moscow, Russia

[2] Moscow State University, Moscow, Russia

toldova@yandex.ru, max.ionov@gmail.com

Abstract. Coreference resolution task is a well-known NLP application that was proven helpful for all high-level NLP applications: machine translation, summarization, and others. Mention detection is the sub-task of detecting the discourse status of each noun phrase, classifying it as a discourse-new, singleton (mentioned only once) or discourse-old occurrence. It has been shown that this task applied to a coreference resolution system may increase its overall performance. So, we decided to adapt current approaches for English language into Russian. We present some quality results of experiments regarding classifiers for mention detection and their application into the coreference resolution task in Russian languages.

Keywords. Coreference resolution, discourse-new detection, singleton detection, discourse processing, natural language processing, machine learning.

1 Introduction

Coreference resolution is the task of grouping mentions into clusters corresponding to subjects (referents). It is an important task for a number of high-level NLP applications, such as machine translation, summarization, and storyline detection. It has been the subject of the lot of research in last 3 decades. However, it is still an open problem. One possible improvement consists on integrating a mention detection module.

Referents receive a different level of attention in the discourse: some may appear only once, others reveal themselves through the discourse. In other words, mentions have different lifespans ([29]). Mentions that appear only once are called *singletons*. By definition, those referents cannot be coreferent. Filtering non-coreferential noun phrases (**NPs**) can improve the coreference resolution results. Besides singletons, it can be useful to detect NPs that introduce new referents that are repeated further in discourse (**DN**, discourse-new mentions) and to differentiate these NPs from recurrent mentions (**DO**, discourse-old). It was discussed in the literature (e.g. [15, 25]) that the DN detection can improve the quality of coreference resolution. 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.

For languages with overt articles, like English, we must decide whether an NP introduces a new referent in spite of an overt definite marker. There are quite a number of papers investigating the impact of different features for this task.

In Russian, which is the article-less language, the task is more complicated. There are no special grammatical clues for detecting NPs referring to new vs. old information. Moreover, for many NP types there are three possible interpretations (besides first mention and repeated mention). An NP could also have a non-specific generic or predicative function (see 3.1 for details). However, there is some theoretical research on reference maintenance that shown that certain features are useful for detecting first mentions

of discourse-salient referents (e.g. [28, 7, 18]). For instance, there is a tendency for introductory NPs to be longer in average 3.3. These NPs usually have qualitative adjectival modifiers. There are also special article-like lexical clues such as 'another', 'new', 'one more' that serve to mark non-identity of an NP referent to previously mentioned ones (see 4.3). There are also lexical features useful for singleton detection, these are different kinds of indefinite and negative pronouns.

There is no comprehensive discussion of the DN detection technique in coreference resolution for Russian language in the literature. Thus, we, firstly, try to set off the possible features for DN detection in Russian texts. We tested features used in English-oriented systems. We wanted to know if some of them are useful for Russian. We also examine theoretical assumptions concerning the DN descriptions properties in Russian (c.f. [2, 33, 5] among others) from the perspective of being a source for DN detection features. On this basis, we suggest an overview of linguistic means that can be used as markers of DN mentions discussed in the literature.

Next, we describe two experiments on discourse status detection: one for singleton detection and another for discourse-new detection. We trained classifiers to detect both kinds of mentions using different features. We show that the features we employ are adequate for the task and produce satisfactory results.

Finally, we provide two experiments on incorporating the mention detection into a coreference resolution system for Russian.

To sum up, we examine features that serve as discourse-new and singleton detectors and show how they improve coreference resolution.

The rest of the paper is structured as follows. Section 2 describes the theoretical grounds for our experiments. Section 3 describes the selected approaches for discourse status detection in Russian as an article-less language. Section 4 describes our experiments. Subsection 4.1 describes the data used for the experiments. In 4.2 we describe the experiment of singleton detection. The experiment of discourse new detection is described in section 4.3. Section 4.4 is devoted to applying the discourse status detection to the coreference resolution task using two approaches: filtering the singletons (subsection 4.4.1) and using the detectors as features for the main classifier (subsection 4.4.2). Section 5 concludes our paper.

2 Background

2.1 Methods for Coreference Resolution

The detection of coreference relations between NPs is a detection of all mentions of the same entity through the text. Consider the following example:

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

In (1), the three co-indexed NPs refer to the entity 'professor Vagner'. These are the proper name *Vagner*, the title of Vagner's occupation *the professor* and the anaphoric pronoun *him*.

Most applications use various machine-learning techniques to get the resulting coreference chain. One basic approach consists of creating a set of pairs of noun phrases (e.g. <*I, Vagner*>, <*Vagner, the professor*>, <*I, the professor*>, etc.), and create a classifier that can predict whether a pair is coreferential one or not (cf. [14], [21], [26] etc.). Baseline systems use different formal features such as token distance, morphological congruency, syntactic features etc. (e.g. Hobbs' syntax-based anaphora resolution algorithm ([12]). Recent systems take into consideration the non-coreferential singletons as well. We use a similar approach to those of the baseline systems (see 3.5 for further details).

2.1.1 Overview of Discourse-new Detection Algorithms

The majority of works about discourse-new detection deal with English texts. Poesio et al. ([25]) presents one of the most thorough analysis of the discourse-new definite (DN) descriptions detection algorithms. The following discussion is based on this work.

It was believed that definite descriptions refer to entities mentioned in the previous discourse. However, nearly 50% of definite descriptions in a

text are discourse new (as shown in [28]and [38]). Consider the following example:

(2) *Google's latest autonomous car is truly driverless, meaning the driver is free to take his hands off the wheel and maybe even text or read a book.*

In this case, the first sentence in an article is a definite NP *the driver* which refers to a driver mentioned for the first time. Such definite expressions can influence the accuracy of coreference chains detection. One of the ways to improve the accuracy is adding a component for detecting discourse-new descriptions (e.g. [38]) into the coreference resolution system. Thus, three questions arise:

1. Which are useful heuristics or features for this component?
2. What is the best scheme for integrating the component into the general coreference recognition process?
3. How much improvement produces to the overall coreference resolution system performance?

Bean and Riloff's system for identifying discourse-new definite descriptions ([4]) is one of the earliest [25]. They suggest the unsupervised method for DN feature collection based on the following heuristics:

1. First sentence extraction heuristic: an NP extracted from the first sentence of a text is discourse-new.
2. Pattern extraction heuristic: a more general pattern can be extracted from the DDs found in the first sentence using the existential head pattern method (e.g. 'N + Head noun' extracted from 'N + Government' from *the Salvadoran Government* and *the Guatemalan Government*).
3. Definite only descriptions heuristic: extracting NPs with high definite probability (e.g. the National Guard).

Special lexemes that serve as introductory markers could be extracted with the previous approach. Such lexical lists can be helpful for article-less languages 3.34.3. Another early approach was an algorithm proposed by Vieira and Poesio ([38]). Earlier ([27]), authors found out that 52% of DDs are discourse new. After that, they proposed to incorporate a set of heuristics for detecting discourse-new descriptions into the algorithm for definite description resolution. Their algorithm identifies five categories of definite descriptions licensed to occur as first mentions on semantic or pragmatic grounds:

1. Semantically functional descriptions" ([20]) such as *the best* or *the first*.
2. Descriptions serving as disguised proper names such as *The Federal Communications Commission*.
3. Predicative descriptions, including appositives and NPs in certain copular constructions, such as *Mr. Smith* or *the president of . . .*
4. Descriptions established (i.e., turned into functions in context) by restrictive modification -particularly by establishing relative clauses ([20]) and prepositional phrases as in *[The hotel where we stayed last night] was pretty good*.
5. Larger situation definite descriptions ([10]) which denote uniquely on the grounds of shared knowledge about the situation (Löbner's 'situational functions'), i.e., definite descriptions like *the sun*, *the pope*, etc..

This classifier had to split definite descriptions into three classes: anaphoric, bridging or discourse new descriptions.

Ng and Cardie ([22]) suggest a set of 37 features for this task (see 2.1.2 for further details). They include their discourse new detector into a coreference resolution system although the authors report no improvement. However, further testing has shown that the way of combining DN detection with the basic coreference resolution module matters (see [25, 15] for details). Other approaches examined in [25] were proposed

by Bean & Riloff ([4]), Poesio and Alexandrov Kabadjov ([24]), Uryupina ([37]). As for the latter approach, author trained two separate classifiers (a DN detector and a uniqueness detector) and used NP's definiteness probability[1]. There are also some more recent approaches based on the tf.idf weight of an NP n-grams suggested in ([30]).

Kabadjov ([15]) thoroughly tested the DN detection module contribution into coreference resolution systems. His experiments shown a significant improvement in performance.

In [15] the *GuiTAR* system is suggested. The whole procedure consists of two processes:

1. Construction of a discourse model.
2. Anaphors resolution.

The ongoing discourse model is used to interpret new NPs. NPs introduce forward-looking centers (see [8]), which means that the system tries to find an appropriate antecedent in both directions.

2.1.2 Ng and Cardie Approach

Most of DN detection techniques are based on set of the features described in [22] with small extensions. The authors suggest the following features:

(a) lexical features: features telling if the target NP and its head overlap with a previous NP including its head, e.g. 'head-match'

(b) grammatical type features: features concerning the 'determiner-like' types of NP modification such as particular types of articles, pronouns or quantifiers (e.g. 'demonstrative', 'possessive' etc.). Also, there is a feature that tells is the absence of any modifier of one of these types

(c) properties and relationships features: this group includes binary features depending on whether the target NP occupies a certain position in some special types of constructions, e.g. is a first part of an appositive construction, or contains a proper noun, or is premodified by superlative etc.

(d) syntactic pattern features: those features refer to more detailed syntactic patterns of the target NP (e.g. noun+Proper Noun, Adjective + noun etc.)

(e) semantic features: this type is about checking some semantic types of the target NP or relations between the target NP and a preceding one (e.g. WordNet relation, etc.)

(f) positional features: these features check NP position in current text, e.g. is the target NP is in the first sentence of a text or in the header of the text?

Those features are also valuable for coreference resolution in article-less languages; they can be adopted for the corresponding systems with some modifications (see sections 3 and 4).

3 Prerequisites for Discourse Status Detection in Russian

3.1 Data Analysis

Languages without articles, such as Russian, do not have specialized grammatical devices for marking a newly introduced referent. Consequently, an NP referring to the first mention of an entity in a particular text can be erroneously attributed to a coreference chain for another entity of the same taxonomy class mentioned earlier in the discourse. We would refer to an NP without determiners and other formal markers of definiteness such as demonstratives or possessive pronouns as to a bare NP. Thus, the referential conflict for an NP in a text is more complicated than in languages with articles. Consider the following example:

(3) *Petrov sozdal [kompaniju] v 2015 godu.*
   a. "*Lit.* Petrov established [company] in 2015."
   b. "In the year 2015 Petrov established [a company]"
   c. "[The company] was established by Petrov in 2015"

[1] It could help to detect unique referent NPs such as *the sun*, *the Urals*, etc.

d. Petrov sozdaet kompaniju kazhdyje tri goda.

'Petrov establishes [a company] every three years'

e. Petrov ne umeet upravlyat' kompanijej.

'Petrov is not able to run [a (any) company]'

In (3) a bare noun *kompaniya* can denote both a before-mentioned entity and a newly introduced one. Some NPs could have more than two possible interpretations:

(a) definite expression referring to a known before mentioned referent (3b),

(b) an indefinite specific NP (referring to a particular newly introduced referent) as interpretation of (3) suggested in (3c), It also can denote an indefinite non-specific NP as in 3d and 3e:

(c) non-referential as in (3d);

(d) generic (3e).

The case becomes more complicated when the descriptor chosen for the first mention of a referent is not the same as in the next mention, as in (4).

(4) *Rabochiye nashli [dva strannyh predmeta]$_i$ na dne transhei, kotoruju oni ryli. [Bronzovye figurki dikogo barana]$_i$ vesili odna — 4.1 kg, drugaya 3.8.*

'The workers found [two curious items]$_i$ at the bottom of a trench they were digging. The bronze [mouflon statues]$_i$ weighted: one was 4.1kg, the other one was 3.8kg'

The NP *bronzovye figurki dikogo barana* has no overt clue for referring to the non-first mention of an entity in Russian in contrast to its English counterpart NP *the bronze mouflon statues* where the definite article indicates the high probability of its antecedent NP presence in the previous text. Another problem is that a generic use of an NP can intervene between the two other identical NPs referring to a specific definite entity of the same taxonomic class. However, there are some clues "signaling" that the referent of the NPs is a newly-introduced entity that would be in focus for a discourse unit longer that a sentence. A general 'classifier' term is used (words like *thing, item* etc,). It was modified with an evaluative adjective *curious* which also served as a marker showing that the entity is in focus of attention (see 3.3 for more details). In this case, the information on discourse structure and the discourse status of a referent might be helpful.

Thus, algorithms elaborated for English could not be used as a ready-made technique for Russian and other article-less languages. Although some of the issues are the same for Russian and for English, the task of discourse-new vs. discourse-old detection should be reformulated for Russian. It concerns the so-called bare NPs' status interpretation: whether they have the generic interpretation, or they are definite specific or indefinite ones. One of the sources of the possible features for DN detection are special introductory markers for discourse salient referents.

3.2 Coreference Models for Referent Tracking in Discourse

As it has been mentioned in 3.1, one way of resolving ambiguous interpretations for article-less languages is to detect the discourse status of NPs. These observations for Russian go in hand with different cognitive-based coreference models as well as typological findings. As it has been shown in [1, 7, 9], the discourse status of a referent imposes the constraints on the feasible NP structural and semantic types (e.g. the preference of anaphoric pronouns for more prominent referents, the "heaviest" NP for a first-time mentioned referent). This hierarchy of referents based on the notion of topic ([7]), or prominence ([28, 1]) corresponds to the hierarchy of different structural types of NPs (from zero anaphora up to full NP). Moreover, more times a referent (an entity) is mentioned in discourse more reduced means to refer to it are used (up to zero pronouns).

There are some models based on the notion of the discourse status suggested and tested for Russian. A. Kibrik (e.g. [16, 17]) worked out the model of a referent activation and tested it for predicting the anaphoric pronoun choice in English

and Russian. In [17], he reports the results of a neural network system based on this model. It is based on the measuring activation status of NP's referent and predicts the choice of a particular NP type in a certain text position.

In [33] the theoretical account for the choice between different kinds of full NPs based on the notion of focus of attention is provided (cf. [9]). The reference maintenance model suggested in [33] is based on a general assumption that referents at a particular point of a text are organized hierarchically. This hierarchy corresponds to the hierarchy of discourse units. Licensing of certain NP types for a referent in a particular point of discourse depends on whether the referent is in a focus of the corresponding discourse unit. In some cases, the speaker could use semantically reduced NPs (a bare noun without any modifier or an anaphoric pronoun), or semantically "expanded" NPs (a noun phrase in which new information is included as in 4). In other contexts the speaker must use special devices to maintain the reference, for instance, the noun has to be modified with a demonstrative pronoun or a special marker of the global focus of attention (e.g. the pronoun *nash* 'our' corresponding to the referent that is the main topic of the discourse). There are linguistic means (lexemes, constructions, word ordering, etc.) that indicate whether the focus of attention in a new discourse span remains the same as in previous one or it has changed. Besides, there are certain linguistic means that serve as signals for the introduction of a salient referent (e.g. ex. (4) in section 3.1). Note that we do not deal in this paper with the ellipsis problem in anaphora resolution [6].

3.3 Features for Introductory NPs

According to the accessibility hierarchy, it is highly unlikely that the discourse new description would be a zero anaphor or a semantically reduced anaphoric pronoun. The cataphoric use of anaphoric pronouns is quite rare. Thus, the task of newly introduced descriptions detection concerns the full NPs resolution.

Arutyunova ([2]) describes the different features of full NP descriptions and analyzes them in terms of different discourse functions. The main properties for the first-mention NPs specified by Arutyunova are as follows: length of NPs, number of adjectives higher than average and semantics of adjectives. She also mentions a special predicate types for the referent introduction such as existential predications (c.f. features for discourse new descriptions detection suggested by Ng & and Cardie ([22]). These observations are summed up in [33] and also discussed in [5] where the corpus analysis of introductory NPs in a special kind of mass media texts is presented.

These papers suggest a list of first mention NPs features for Russian (some of them coincide with the above-discussed features for English).

A. Introductory NPs tend to occur in the focus part of the utterance. In other words, there is a tendency for such NPs to occupy the position closer to the end of the sentence.

B. There are specific existential or quasi-existential constructions introducing a new referent into the discourse. Such as the sentences with the verbs of a referent existence causation such as *vozniknut'* 'to emerge', *poyavit'sya* 'to appear', *sozdat'* 'to create' and many others (for a more detailed list see [5] (cf. the constructional features in [22])

C. Length of introductory NPs statistically significantly differs from the average length. Cf. table 1.

D. The number of pre-modified adjectives is higher for introductory NPs relative to the average number of premodified adjectives

E. There is a tendency to include non-relational evaluative adjectives into introductory NPs. Besides, there is a tendency to include additional so-called encyclopedic or factual information (c.f. the tendency to use the expression 'x-year old' in the first-mention NP for a not well-known person in English news reports).

F. There is a special NP type so-called under-specified NP that is used to mark highly salient referents. That is an NP with a unspecific

classifier such as 'item', 'building', 'creature', 'figure', 'construction' etc. as a Head noun and with an evaluative adjectives such as 'mysterious', 'strange', 'curious', 'nice', etc. as modifiers (c.f. curious items used to refer to statues in 4).

G. There is a special class of 'alternators' 'signaling' the inequity of the NPs referents. There are several classes of such alternators:

   (a) indefinite markers *odin* 'one', *nekij* 'a person';

   (b) inequity markers such as *drugoj, inoj* 'other', etc.;

   (c) similarity markers such as *takoj* 'such, of this kind', *podobnyj* 'analogous', *pohozhij* 'similar', etc.;

   (d) markers that introduce an element of a set *odin iz* 'one of the';

   (e) *ostal'nie* 'the rest';

   (f) the order of introduction: *pervyj iz (nih)* 'the first', *vtoroj* 'the second', *poslednij* 'the last'.

   Although these alternators are not very frequent in discourse, they are reliable features for the discourse-new detection.

Table 1. Average discourse new NPs length in comparison to the average NPs length in coreference corpus for Russian

	Full NPs	Disc-new	Disc-old
Mean	1.909	2.951	1.668
Std dev.	1.753	2.620	1.378

3.4 Features for Singleton Mentions

Features used for singletons detection are the same as for discourse-new detection. Thus, while the NP non-repetition or the head of an NP non-repetition in the previous context are relevant features for detecting both mentions classes, the unique NP or the unique head is much more likely for the former ones. In [35], 4 groups of features are tested for singletons detection: basic, structural, lexical, and (quasi-)syntactic features. Most of the features were proposed before for detecting singleton mentions in English (e.g. [29, 22]). Some other features, correlated with entity discourse role, were used in the first mention detection task (see also [35]). Thus, the set of features for DN detection should combine features for detecting non-anaphoricity with those that should have a correlation with the discourse role: non-coreferent mentions should be less important for the discourse.

As has been mentioned above the syntactic role is one of the important features for detecting all the three mentions classes. However, the non-argument NP position can also play a role. In this research, we use the noun case as a correlate for a syntactic role (cf. nominative case for Subject vs. Accusative for Object vs. others). We also employed genitive/non-genitive case as a separate feature. The source for this feature was the intuition about Russian genitive that it coincides with non-argument positions.

There are also some special lexical features for singletons detection. There are special indefinite pronouns classes, namely non-specific pronouns (e.g. *chto-nibud'* 'something'), free-choice pronouns (e.g. *ljuboj* 'any'), distributive quantifiers such as *kazhduj* 'every', and negative pronouns. These NPs are non-referential so usually, they are unable to denote repetitive discourse entities.

3.5 A Baseline Method for Coreference Resolution

In order to show the impact of discourse status detection for the coreference resolution task we created a simple baseline coreference resolution system for Russian. To do so, we reproduced the system described in [34]. The method described there was based on an approach proposed by Soon et al. ([32]), a basic ML approach widely used as a baseline for various languages. According to this approach coreference chains are formed from the pairs of coreferent noun phrases.

The system uses several types of features: string similarity, morphological features, lexical, basic syntactic and very basic semantic features. More

specifically, the feature set consists of N features. The feature set is fairly standard. It includes:

1. String match: tells if noun phrases are the same or one is an acronym of another.
2. Morphological agreement: number, gender, properness, animacity.
3. Morphological features: types of pronouns if the NPs are pronouns.
4. Semantic agreement: tells if two NPs are named entity of the same class or one is an alias of another.
5. Two noun phrases are in the appositive relation.

The quality of the system is presented in the table 2.

Table 2. Baseline coreference resolution system performance

	MUC			$\mathbf{B}^3$		
	P	R	F_1	P	R	F_1
Baseline system	40.47	52.88	45.85	25.76	40.93	31.62

4 Experiments

To check how the features proposed in the previous sections allow us to detect the discourse status of a noun phrase, we built a set of classifiers with different sets of features both for the task of singleton detection and first-mention detection, and analyzed the quality of these classifiers and their impact on the task of coreference resolution.

Before describing the experiments we should describe the corpus that was used for training and testing the classifiers.

4.1 Data

Our experiments were conducted on RuCoref, a corpus of Russian texts with coreference annotation[2] released during RU-EVAL evaluation forum ([36]).

This corpus consists of short texts in a variety of genres: news, scientific articles, blog posts and fiction. The whole corpus contains about 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 ([31]). Noun phrases were obtained using a simple rule-based chunker ([13]). The corpus was randomly split into train and test sets (70% and 30% respectively).

Since the RuCor annotation followed MUC guidelines ([11]), singletons are not annotated in the corpus, so every unannotated noun phrase was considered a singleton. This means that we do not distinguish mentions that are never coreferent and potentially coreferent mentions used only once in a text, even though they may have, in principle, very different structure.

The dataset is highly unbalanced: recurring mentions, first mentions and singletons are in the ratio 1:4:40. To overcome this problem, we performed a sampling on the training set for training both detectors. The best results were achieved using the combination of oversampling and undersampling methods ([3]) and was implemented in the imbalanced-dataset Python module ([19]).

4.2 Singleton Detection

We are using 4 groups of features for this experiment: basic, structural, lexical, and (quasi)syntactic features. Most of the features we used were proposed before for detecting singleton mentions in English texts (e.g. [29, 22]). Some other features correlated with entity discourse role are also used in the first mention detection task (section 4.3, see also [35]).

[2]The corpus may be freely downloaded at `http://rucoref.maimbava.net/`.

As it was already mentioned, our notion of singletons combines two types of mentions: those that can not be anaphoric and those that could be anaphoric but were mentioned only once in a discourse fragment. In order to detect both groups, we compiled features that detect non-anaphoricity as well as those that should be correlated with a discourse role: non-coreferent (i.e. singleton) mentions should be less important for discourse and have lower discourse role.

4.2.1 Basic Features

The most basic feature is the number of occurrences of a candidate NP or its head in a text before. It is obvious that if an NP is repeated, chances are, this is the same mention and hence the entity is not a singleton.

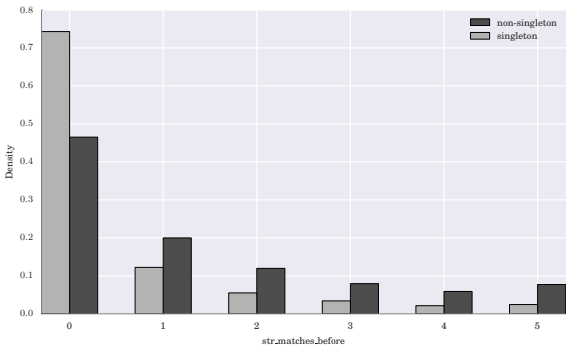
The distribution of those features over a train set confirms this idea showing a significant difference for two target classes (see figure 1). Other features from this group include binary flags like whether a noun phrase is animate, a proper noun, contains non-Cyrillic characters or is a pronoun. Some of those features were shown to be useful for English (e.g. [29]).

4.2.2 Structural Features

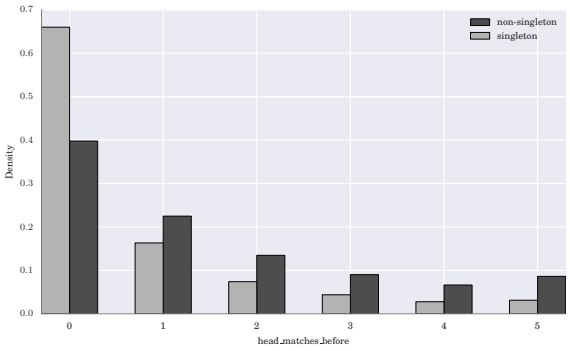
This group contains two features: NP length in words and the number of adjectives in the NP before its head. They both correlate with an entity importance in the discourse: the more important an entity is, the more words would be spent on it. These two features has theoretical motivation and showed great impact in the first mention detection task ([35]), showing their correlation with a discourse role of a mention. Figure 2 shows the distribution of these features over a train set.

4.2.3 Quasi-syntactic Features

Syntactic structure can shed the light on the discourse role of a noun phrase. Studies in the Centering theory and various other discourse studies showed that coreferent mentions tend to be core verbal arguments and prefer sentence-initial positions in a sentence (e.g. [9, 39]).



(a) A number of occurrences of a full NP

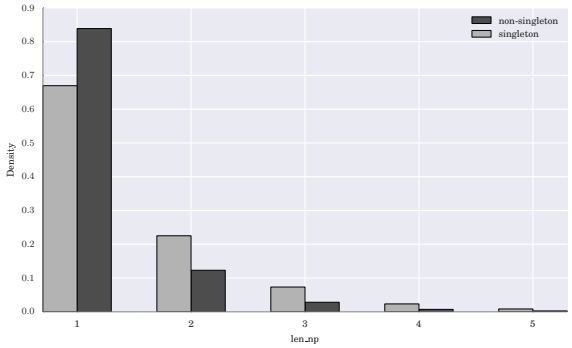


(b) A number of occurrences of the head of an NP

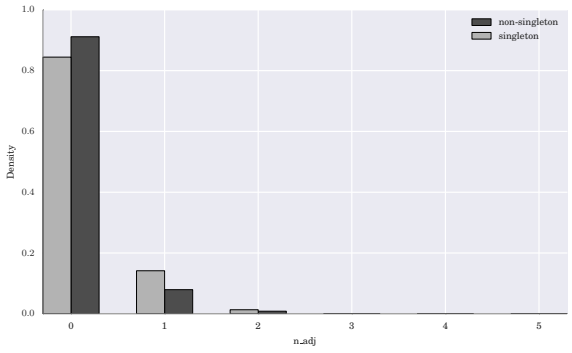
Fig. 1. A number of occurrences of a candidate NP in a previous discourse

Since there is no available reliable way to automatically annotate verbal arguments for Russian, we used heuristics instead that worked in the following way: if an NP is either in Nominative case or in the beginning of the sentence, we thought of it as a subject, if an NP is both in Accusative case and in the end of the sentence, it was considered an object. While the first heuristic performed well, the second yielded too many mistakes (partly because of mistakes in morphological annotation), so it was not present in the final feature set.

A language-specific and less-standard feature that we employed was if an NP is in the Genitive case. The source for this feature was an intuition about Russian Genitive that coincides with non-argument positions. Judging from the distribution of this feature over the training set



(a) A number of words in an NP



(b) A number of adjectives in an NP

Fig. 2. A distribution of the structural features

(see fig. 3) it is clear that there is a correlation but not as strong as for the previous features.

4.2.4 Lexical Features

While all previously described features were designed to detect mentions that are not important enough for the discourse to be mentioned more than one time, lexical features were designed to detect non-anaphoric noun phrases.

For this we used four manually compiled lists of different classes of pronouns: (i) indefinite pronouns, (ii) possessive pronouns and (iii) negative pronouns. These groups are known for the tendency to be non-referential, therefore the presence of such lexical markers can be used to detect singletons with high degree of confidence.

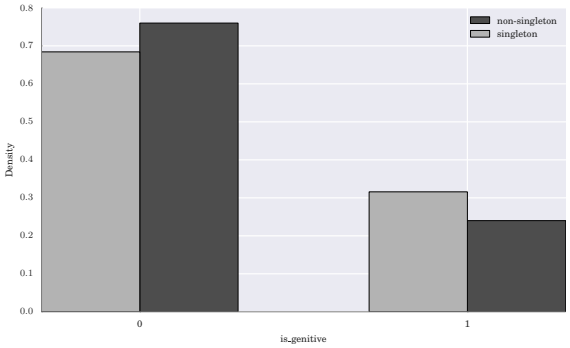


Fig. 3. A Genitive case of an NP

4.2.5 Results

To test how good various groups of features distinguish singleton mentions from non-singleton ones we have built a set of classifiers. As a baseline we used a simple heuristic: an NP was considered a singleton mention if and only if there was no such NP or its head before. To implement the classifier, we used a Random Forest classifier from the scikit-learn Python library ([23]). Results of the experiments are presented in the table 3.

Table 3. Singleton classification results (for the minority class)

	P	**R**	**F1**
Baseline	0.423	0.659	0.515
Basic	0.463	0.736	0.569
Basic + Struct	0.473	0.740	0.577
Basic + Struct + Lists	0.493	0.744	0.593
All features	0.499	0.736	0.595

Results are far from perfect but even the most basic feature set performs better than the baseline. Adding more sophisticated features further improves quality.

4.3 First Mention Detection

We trained a classifier to distinguish discourse-new from discourse-old mentions. As it was shown before, those two classes of mentions are structurally different, which means that it is possible to use structural features to distinguish them.

Singleton noun phrases poses a problem for the experiment: on the one hand, they appear in the discourse for the first time hence they are by definition discourse-new. On the other hand, the fact that referents they represent appear just once means that they are less important for the discourse than other referents. So their structure should differ from the structure of the noun phrases that introduce a new referent that is salient for the discourse. Figures in section 4.2 support this hypothesis. So, in order to decrease noise in our data we used only non-singleton mentions for this experiment.

We used 3 groups of features to distinguish discourse-new from discourse-old mentions: (a) basic features like the number of occurrences of the noun phrase in the previous discourse, (b) structural features, and (c) lexical features.

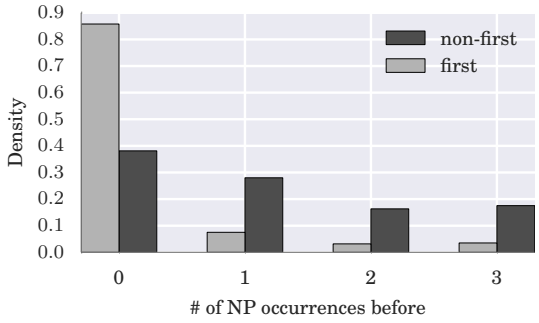
Features that we used in this experiments are mostly the same as in the previous one since in both experiments noun phrases differ in their discourse status and the features are designed to detect it.

4.3.1 Basic Features

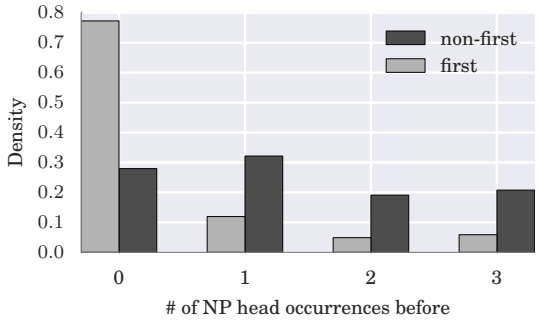
The basic feature set for this experiment is the same as in the previous one. It includes the number of occurrences of the noun phrase and its head in the previous discourse. Figures 4 shows the distribution of those features over the train set. Other features in this group include some properties features of the noun phrase that correlates with its discourse status: whether it is a proper noun, consists of uppercase characters or contains Latin symbols.

4.3.2 Structural Features

Again, as in the previous experiment, this group contains two structural features: length of NP in words and the number of adjectives in the NP before its head. They both correlate with an entity importance in the discourse: the more important an entity is, the more words would be spent on it. Figure 5 shows the distribution of these features over a train set.



(a) A number of occurrences of a full NP



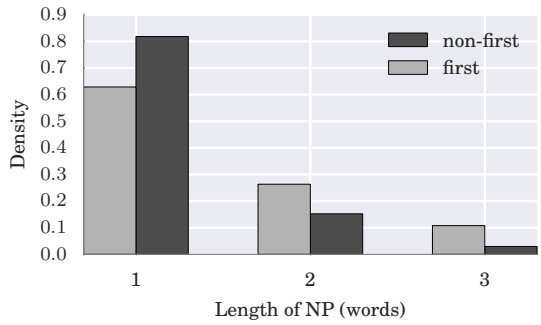
(b) A number of occurrences of the head of an NP

Fig. 4. A number of occurrences of a candidate NP in a previous discourse

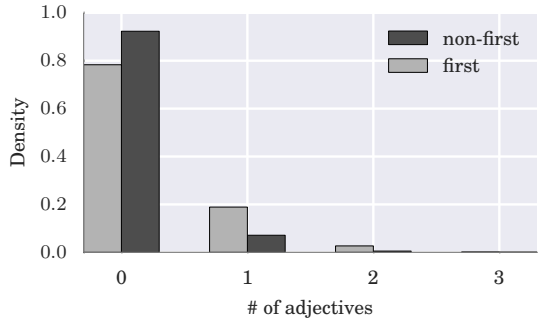
4.3.3 Lexical Features

As has been shown in section 3.3, there are special lexical aids that introduce new referents in the discourse, alternators. The presence of such markers in the noun phrase indicates that this NP is a discourse-new mention. We used 6 manually created lists of such markers:

1. General class names: nouns that define a class (*building*, *manager*, etc.);
2. New referent introductory adjectives (*contemporary*, *latest*, etc.);
3. Non-identity and similarity markers: *another*, *similar*, etc.;
4. Common knowledge markers (*famous*, *legendary*, etc.);



(a) A number of words in an NP



(b) A number of adjectives in an NP

Fig. 5. A distribution of the structural features

5. Adjective markers of a discourse role in an NP (*main*, *small*, etc.);
6. Subjective markers (*good*, *prestigious*, etc.);

Additionally, we used lists of possessive, demonstrative and indefinite pronouns as markers of a discourse status.

We extracted a list of adjectives that are most important for the classification in order to increase coverage of lexical features. To do so, we performed univariate feature selection operation using 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 have manually cleaned this list. We removed the pronouns and words erroneously tagged as adjectives. From the cleaned list we extracted 50 most important adjectives.

Top 10 adjectives from the list are presented in the table 4.

Table 4. Top 10 adjectives most valuable for classification

#	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

4.3.4 Results

We used a Random Forest classifier from the scikit-learn Python library ([23]). 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 5. We report precision, recall, and F1-measure for each feature set.

All feature sets, including the lexical lists, increase precision at the cost of recall as shown in Table 5. The combination of all features shows the best results. There are several ways for further improvement: (a) reducing noise in the data (e.g. chunker used to find noun phrases in the data can not handle complicated noun phrases therefore structural features are not precise), (b) improving lexical features manually and automatically, (c) adding more sophisticated

Table 5. First mention classification results (for the minority class)

	P	R	F1
Baseline	0.526	0.830	0.644
String	0.533	0.827	0.649
String + Struct	0.548	0.806	0.653
String + Struct + Lists	0.560	0.796	0.658

ISSN 2007-9737

features (e.g. whether a noun phrase is an apposition).

4.4 Applying the Discourse Status Detectors to the Coreference Resolution Task

We apply the discourse status detectors described in the previous sections to the baseline coreference system. We tried two ways of applying them: as a separate preprocessing step, and using the output of those classifiers as features of the mention-pair classifier.

4.4.1 Filtering a List of Candidates Using Discourse Status Detectors

The first approach was to use the detected discourse status in the preprocessing step to filter the list of NP pairs, removing those pairs that contained detected singletons. We used the singleton detection on every possible candidate pair. If the probability of being a singleton was above the threshold, the pair was discarded. Results with different thresholds are presented in the table 6.

Table 6. Coreference resolution with singleton filtering

	MUC			$\mathbf{B}^3$		
	P	R	F_1	P	R	F_1
No filter	40.47	52.88	45.85	25.76	40.93	31.62
Thresh=0.1	43.54	50.13	46.60	27.60	37.55	31.82
Thresh=0.2	43.52	49.78	46.44	27.49	37.07	31.57

Even though the recall has decreased due to filtering some false negatives, we can see that the precision of the system with filtered singletons is better than the precision without mention detection. On the other hand, increasing the threshold lowers the quality making the applicability of this method very limited. However, the singleton detector quality is quite low ($F_1 = 0.595$, see table 3) and needs further improvement.

4.4.2 Using Discourse Status Detectors as a Features for a Mention-pair Classifier

The second approach is to use the discourse status detected using the classifiers discussed above as a feature for the coreference classifier. We tried three different setups: (a) a baseline classifier plus a feature with a result from a discourse-new classifier, (b) a baseline classifier plus a feature with a result from a singleton classifier, and (c) a baseline classifier plus both features. Results are given in the table 7.

Table 7. Coreference resolution with mention detection used as features

	MUC			$\mathbf{B}^3$		
	P	R	F_1	P	R	F_1
No filter	40.47	52.88	45.85	25.76	40.93	31.62
Singletons	41.89	50.62	45.84	27.66	39.41	32.51
DN	45.09	51.40	48.04	27.10	39.55	32.16
Both	42.39	48.97	45.44	27.30	38.11	31.81

Table 7 shows that each feature improves the quality of the coreference resolution. The discourse-new detection improves the MUC-score dramatically increasing the precision, which means that this detector is useful to cut long erroneous chains if one of the mentions is discourse-new. Detecting singletons increases the precision while decreasing the recall for both metrics. This means that this feature helps filtering some false positive pairs, but at the same time it filters some true positives.

Using both features at the same time gives an unexpected decrease in performance. The precision of this setup is still higher than the precision of a baseline system, but the recall is significantly lower and the overall quality is lower than when using features individually. This result requires further investigation and probably these detectors should be applied in a more sophisticated way.

5 Conclusions

We described an approach for creating two discourse status detectors in this paper: a

singleton detector and a first mention detector, using structural theoretically motivated features and manually and semi-automatically created lists of lexical markers. We showed that these detectors can improve the quality of coreference resolution for an article-less language.

The impact of those detectors on the coreference resolution quality may be further improved by improving the quality of the detectors by using more sophisticated features and improving the features that we used.

Theoretically motivated lexical features shows promising results and further investigation of this type of features should improve the quality of the discourse status detection task and, as a result, the overall quality of coreference resolution.

Acknowledgments

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

References

1. **Ariel, M. (1990).** *Accessing Noun-Phrase Antecedents*. Routledge.
2. **Arutyunova, N. (1980).** Nomination, reference, meaning. [nominaciya, referenciya, znacheniye] (in Russian). In *Nomination: General Questions. [Nominaciya: obshie voprosi]*. Nauka.
3. **Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003).** Balancing training data for automated annotation of keywords: a case study. *WOB*, pp. 10–18.
4. **Bean, D. L. & Riloff, E. (1999).** Corpus-based identification of non-anaphoric noun phrases. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 373–380.
5. **Bonch-Osmolovskaya, A., Toldova, S., & Klintsov, V. (2012).** Introductory noun phrases: a case of mass media texts. [strategii introduktivnoj nominacii v tekstah smi] (in Russian).
6. **Gelbukh, A., Sidorov, G., & Bolshakov, I. (2002).** On coherence maintenance in human-machine dialogue with contextual ellipses. *Computación y Sistemas*, Vol. 5, No. 3, pp. 204–214.
7. **Givón, T.**, editor **(1983).** *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamins., Amsterdam.
8. **Grosz, B. J. & Sidner, C. L. (1986).** Attention, intentions, and the structure of discourse. *Comput. Linguist.*, Vol. 12, No. 3, pp. 175–204.
9. **Grosz, B. J., Weinstein, S., & Joshi, A. K. (1995).** Centering: A framework for modeling the local coherence of discourse. *Comput. Linguist.*, Vol. 21, No. 2, pp. 203–225.
10. **Hawkins, J. A. (1978).** Definiteness and indefmiteness: a study in reference and grammaticality prediction. *London: Croom Helm*.
11. **Hirschman, L. & Chinchor, N. (1998).** Appendix f: Muc-7 coreference task definition (version 3.0). *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.
12. **Hobbs, J. (1978).** Pronoun resolution. *Lingua*, Vol. 44, pp. 339–352.
13. **Ionov, M. & Kutuzov, A. (2014).** Influence of morphology processing quality on automated anaphora resolution for Russian. *Proceedings of the international conference Dialogue-2014*, RGGU.
14. **Jurafsky, D. & Martin, J. H. (2009).** *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
15. **Kabadjov, M. A. (2007).** *A comprehensive evaluation of anaphora resolution and discourse-new classification*. Ph.D. thesis, Citeseer.
16. **Kibrik, A. (1983).** Ob anafore, dejksise i ix sootnoshenii [on anaphora, deixis, and the correlation between them]. *Razrabotka i primenenie lingvisticheskix processorov (ed. A.S.Narin'jani)*, Novosibirsk, VC SO AN SSSR, pp. 107–129.
17. **Kibrik, A., Linnik, A., G., D., & Khudyakova, M. (2012).** Optimizacija modeli referencial'nogo vybora, osnovannoj na mashinnom obuchenii [optimization of a model of referential choice, based on machine learning]. *Computational Linguistics and Intellectual Technologies*, volume 11, Moscow, RGGU, pp. 237–246.
18. **Kibrik, A. A. (2011).** *Reference in discourse*. Oxford University Press.

19. **Lemaître, G., Nogueira, F., & Aridas, C. K. (2016).** Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *CoRR*, Vol. abs/1609.06570.

20. **Löbner, S. (1985).** Definites. *Journal of semantics*, Vol. 4, No. 4, pp. 279–326.

21. **Mitkov, R. (1999).** *Anaphora resolution: the state of the art*.

22. **Ng, V. & Cardie, C. (2002).** Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 1–7.

23. **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. (2011).** Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830.

24. **Poesio, M. & Kabadjov, M. A. (2004).** A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. *Proceeding of LREC*, pp. 663–666.

25. **Poesio, M., Kabadjov, M. A., Vieira, R., Goulart, R., & Uryupina, O. (2005).** Does discourse-new detection help definite description resolution. *Proceedings of the Sixth International Workshop on Computational Semantics, Tillburg*.

26. **Poesio, M., Ponzetto, S. P., & Versley, Y. (2010).** Computational models of anaphora resolution: A survey.

27. **Poesio, M. & Vieira, R. (1998).** A corpus-based investigation of definite description use. *Comput. Linguist.*, Vol. 24, No. 2, pp. 183–216.

28. **Prince, E. F. (1992).** The zpg letter: Subjects, definiteness, and information-status. *Discourse description: diverse analyses of a fund raising text*, pp. 295–325.

29. **Recasens, M., de Marneffe, M.-C., & Potts, C. (2013).** The life and death of discourse entities: Identifying singleton mentions. *Human Language Technologies: The 2013 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, PA, pp. 627–633.

30. **Ritz, J. (2010).** Using tf-idf-related measures for determining the anaphoricity of noun phrases. **Pinkal, M., Rehbein, I., im Walde, S. S., & Storrer, A.**, editors, *Semantic Approaches in Natural Language Processing: Proceedings of the 10th Conference on Natural Language Processing, KONVENS 2010, September 6-8, 2010, Saarland University, Saarbrücken, Germany*, universaar, Universitätsverlag des Saarlandes / Saarland University Press / Presses universitaires de la Sarre, pp. 85–92.

31. **Sharoff, S. & Nivre, J. (2011).** The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. *Proc. Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo.

32. **Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001).** A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, Vol. 27, No. 4, pp. 521–544.

33. **Toldova, S. (1994).** Focusing and discourse structure as important factors of reference choice in text. [fokus vnimaniya i ierarchija discursa kak vazchnyje factory vybora nominacii ob'ekta v tekste].

34. **Toldova, S. & Ionov, M. (in press).** Coreference resolution for Russian: Establishing the baseline.

35. **Toldova, S. & Ionov, M. (in press).** Features for discourse-new referent detection in Russian. *Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016, Konya, Turkey, April 3-9, 2016, Proceedings*, Konya, Turkey.

36. **Toldova, S., Rojtberg, A., Ladygina, A., Vasilyeva, M., Azerkovich, I., Kurzukov, M., Ivanova, A., Nedoluzhko, A., & Grishina, J. (2014).** RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. *Computational Linguistics and Intellectual Technologies*, Vol. 13 (20), pp. 681–694.

37. **Uryupina, O. (2003).** High-precision identification of discourse new and unique noun phrases. *ACL Atudent Workshop*, Sapporo.

38. **Vieira, R. & Poesio, M. (2000).** An empirically based system for processing definite descriptions. *Comput. Linguist.*, Vol. 26, No. 4, pp. 539–593.

39. **Ward, G. & Birner, B. (2004).** Information structure and non-canonical syntax. *The handbook of pragmatics*, pp. 153–174.

Svetlana Toldova is an associate professor of Natural Language Processing at National research university “Higher school of economics”. She obtained her PhD from Lomonosov Moscow State University. Her research areas include Discourse processing, Anaphora and Coreference resolution, Corpus linguistics, Information Extraction, various aspects of Natural language processing. She has been one of the organizers for the Forum of NLP systems evaluation for Russian (morphological tagging, dependency parsing, anaphora and coreference resolution).

Max Ionov is a PhD student at Lomonosov Moscow State University. His research is devoted to Anaphora and Coreference resolution. His other research interests include Discourse processing, Information retrieval, Corpus linguistics and Semantic web technologies.

Article received on 11/10/2016; accepted on 02/11/2016.
Corresponding author is Svetlana Toldova.