Identification of singleton mentions in Russian

Max Ionov[1] and Svetlana Toldova[2]

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

Аннотация This paper describes a pilot study of the problem of detecting singleton mentions in Russian texts. A noun phrase is considered a singleton mention if it is the only referent of some entity.
We discuss various morphosyntactic and lexical features, some of which were used for analogous tasks for English and propose new features derived from the discourse analysis.
Testing the machine learning classifiers trained with the use of proposed features, we conclude that although the quality of classifiers is significantly lower than for English, they still have rather high precision and thus can be helpful in various tasks of mention tracking.

Keywords: coreference resolution, mention detection, discourse processing, natural language processing, machine learning

1 Introduction

Coreference resolution is an important preprocessing step for many advanced NLP tasks that deal with discourse processing. Although this task has been actively researched for several decades (e.g. [11,8]), there is still a substantial amount of work going on in this direction. One of the problems that received much attention in the last years is the improvement of an important preprocessing step — mention detection.

Coreference resolution task can be seen as a task of linking pairs of mentions that refer to the same discourse entity. However it has been shown ([13]) that not every noun phrase in a text likes to be linked, some of them appear just once, henceforth referred to as singletons. Moreover, the fraction of such mentions is very high (about 50%) and, as a consequence of it, the chance of linking a singleton with something by mistake is relatively high. It was demonstrated (ibid) that filtering out singletons before coreference resolution improves its overall quality. In addition, removing a large number of unnecessary noun phrases should improve the computational efficiency of the system. Other high-level tasks that require tracking of discourse mentions can also benefit from singleton detection.

In this paper we conduct a pilot study on singleton detection for Russian. We employ the features that were applied to the similar task for English and

check if they work for Russian. In addition we discuss some new, Russian specific features. In the experiment we test different subsets of them and measure the importance of each feature.

The rest of the paper is structured as follows. In section 2 we give an overview of the previous work on singleton detection and some theoretical literature on discourse that deals with certain peculiarities of singleton mentions. In section 3 we describe the corpus that we used for our experiments, the set up for the experiments and their results. In section 4 we analyze the impact of the features and discuss expected and unexpected outcomes.

2 Background and motivation

2.1 Singleton detection in English

Among the vast amount of research devoted to coreference resolution for English, there has been some work done on various aspects of mention tracking, e.g. discourse-new detection ([18]) — finding NPs that start coreferential chains; distinguishing anaphoric and non-anaphoric NPs ([9]) — detecting definite NPs whose interpretation does not depend on previous mentions. Another closely related type of mention tracking is singleton detection ([13]) — detecting NPs that correspond to the entities that are mentioned only once.

The definiteness category is usually an important factor for those tasks, and some features discussed in the papers rely crucially on the presence of specific articles (see [9], for example). Other researchers employ more complicated features based on determiners. [18] mentions that unique mentions are used with the definite article much more often than with indefinite one, like the singleton "the government" in her corpus was found in 23.9% usages while its "a government" version only in 4.8%. As the result, for every unique noun phrase, the "definiteness probability" was assessed and used as a feature in the further classification.

Since Russian is an article-less language, we could not use features based on the overt morphosyntactic definiteness.

The present work is mainly based on Recasens et al. ([13]) where the impact of different morphosyntactic and semantic features on the singleton vs. repetitive mention detection was analysed. To estimate this impact authors used a binary logistic regression model. The feature set used for it relied on the theoretical assumptions made in the theoretical literature on the referential choice modeling. According to Prince [12], [1] or Centering Theory ([5]), the morphosyntactic properties of a target NP could be relevant. Besides, different types of NPs have different probability of being coreferent. For instance, anaphoric pronouns are very likely to be coreferent, the animacy, number and some other properties of an NP correlate with the likelihood of it being linked to other NPs in the text. According to [13] indefinite or negative pronouns have negative association with chaining.

The other cluster of features discussed there is based on information structure, most importantly, syntactic position. Subjects as well as verb arguments have positive association with coreferent use of an NP.

2.2 Features for non-repeating mentions in Russian

As for Russian, there is a lack of research concerning the unique or non-coreferring expressions properties. However, there are works devoted to special devices for new salient entities introduction into discourse (e.g. [2], [4], [15], and [16] for discourse-new detection).

Among the features used there is NP length (the first NP for a new salient referent tends to be longer) and, in particular, the number of adjectives. For instance, [15] suggests the following examples:

(1) a. ? *On podošol k stolu, vzjal s nego černuju ručku v forme kinžala,* ?*sel i stal pisat'.*

He came up to the table, took a black pen in the shape of a dagger from it ?and started writing.

b. *On podošol k stolu, vzjal ručku, sel i stal pisat'.*

He came up to the table, took a pen and started writing.

In both examples the underlined entity (the pen) is a singleton. In the example 1a it has two modifiers and the end of the sentence without mentioning the entity again sounds unnatural. In the example 1b the NP is a typical ('expected') participant of the 'writing' event. Therefore, the sentence seems acceptable.

According to [16], there are also some specific lexical features, for example, so-called non-identity words like *takoj* (such) and some adjectives have an impact on the status of an NP in the discourse.

Thus, the question is whether those features that showed their usefulness in the first-mentions detection task, could be helpful for the singleton differentiation as well.

3 Experiments

3.1 Data

Our experiments were conducted on RuCor, a Russian coreference corpus released as a part of the RU-EVAL campaign ([17])[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 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 ([14]). For the experiments described in this paper, texts were additionally syntactically parsed using the same tools. The morphological tags were checked and fixed manually, since it was previously shown that errors on this level affects significantly the quality of a related task ([7]). The corpus was randomly split into a training and a test set (70% and 30% respectively).

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

Since the RuCor annotation followed MUC guidelines ([6]), singletons are not annotated, 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, in principle, have different structural properties.

3.2 Features for singleton detection

In this pilot study, we tested 4 groups of features: basic, structural, lexical, and syntactic features. Most of the features that we used were proposed before for detecting singleton mentions in English (e.g. [13,9]). Some other features, correlated with entity discourse role, were previously used in the first mention detection task ([16]).

As we mentioned already, our notion of singletons collapses two types of mentions: those that can not be anaphoric and those that were mentioned only once in the text. In order to detect both of them, we compiled features that detect non-anaphoricity with those that should have correlation with the discourse role: non-coreferent mentions should be less important for the discourse.

Basic features The most basic feature is the number of occurrences of the NP in question or its head in the text before. It is obvious that if an NP is repeated, chances are that the entity is not a singleton. Other basic features check whether the noun phrase is animate, a proper noun, contains non-cyrillic characters or is a pronoun. Those features were shown to be useful for English in [13].

Structural features This group contains two features: NP length (in words) and the number of adjectives. Both of them should correlate with the entity importance in the discourse: the more important an entity is, the more words would be spent on it. Those features showed a great impact on the first mention detection task ([16]), having a strong correlation with the discourse role of a mention.

Syntactic features Syntactic structure can shed a lot of light on the NPs' discourse roles. Studies in Centering theory and various discourse studies showed that coreferent mentions tend to be core verbal arguments and prefer sentence-initial positions in a sentence (e.g. [5,19]).

In order to include this information in the classifier we used the syntactic annotation described in section 3.1. An NP was considered a subject if the type of a syntactic role was 'предик'. An NP was considered an object if its role was '1-компл'[4].

[4] A submitted version of a paper did not use syntactic information because the experiments showed its low quality. Further experiments showed that although there are a lot of errors in the syntactic information, it still improves the classification quality.

Less standard feature that we employed was if an NP is in genitive case. The source for this feature was the intuition about Russian genitive that it coincides with non-argument positions.

Lexical features We used four precompiled lists to detect non-coreferent mentions: (i) indefinites, (ii) possessive pronouns, (iii) negative pronouns and (iv) non–identity words ('takoj zhe' such as, 'similar' podobnyj, etc.). If all the previous groups contained features that were designed to detect NPs that are more likely to be unique in the text, judging by their discourse role, those lists (except for the last one) should detect NPs that can not be coreferent. This way we have features for both non-referring and potentially coreferent but unique mentions.

3.3 Results

To test how good various features distinguish singleton mentions from non-singleton ones, we have built a set of classifiers using Random Forest classifier implemented in Scikit-Learn library ([10])[5]. This classifier was chosen mainly for two reasons: firstly, Decision Tree classifiers and improvements over them are often used in related work. Secondly, using this algorithm provided some insight on feature importances during feature engineering step. As a baseline, we established a heuristic baseline: an NP is considered a singleton mention if and only if there were no such NPs or no NPs with the same head before. Results of the experiments are presented in table 1.

As it has already been noted, singletons are more frequent than coreferent mentions. Due to this disproportion the training set is unbalanced — there are more than two times more singleton mentions than non-singleton ones — and this influences resulting classification quality. To overcome this problem, we performed oversampling on the training set. The best results were achieved using SMOTE+Tomek method ([3]). The results are presented in table 1

	P	**R**	**F1**
Baseline	0.470	0.665	0.551
Basic	**0.621**	0.630	0.626
Basic + Struct	0.601	0.676	0.637
Basic + Struct + Lists	0.609	0.671	0.638
All features	0.600	**0.708**	**0.650**

Таблица 1. Classification results (for the minority class)

The basic feature set gives the highest precision, while the recall is lower than using more complex features. The more features are used, the more is recall. Using the full set of features increases the recall by about 12%.

[5] An IPython notebook with the experiments is available here: https://git.io/vwE9F.

To understand the individual importance of each feature, we have trained a logistic regression model on the training data. The coefficients for each feature are given in table 2. A singleton is the positive class, so positive coefficients mean a bias towards being a singleton mention.

Feature	**Estimate**
Basic features	
str_match_before=0	0.6848
head_match_before=0	0.8824
latin	-0.5225
is_proper_noun	-0.4548
is_animate	-0.8090
is_pronoun	-4.2037
Structural features	
len_np=1	-0.3331
1<len_np<4	-0.0050
len_np>=4	0.1227
n_adj=0	0.1178
n_adj>2	-0.2718
Syntactic features	
is_genitive	0.1040
is_subject	-0.9267
is_object	-0.1767
Lexical features	
in_list_neg_pronouns	1.2797
in_list_non-identity_sim	0.4702
in_list_possessives	-0.1276
in_list_indef	0.6106

Таблица 2. Feature importances

Results for basic features agree with the results reported in [13] for English. The unique head lexeme is a sustainable feature for the unique mentions. Animate and proper nouns, and anaphoric pronouns are more likely to be coreferential.

The NP length (number of modifiers) acts in a less expected way: short and moderately long NPs tend to be non-singletons whereas only very long NPs tend to be singletons. However, this agrees with the result in [13]. Positive number of adjectives correlates with being coreferent as expected.

The syntactic features also have expected behavior: subject mentions tend to be non-singletons and mentions in genitive are slightly biased towards being singletons. This agrees with previously stated intuition of them being non-topical arguments.

Indefinite (free-choice) pronouns such as *lyuboj* 'any', *kazhdyj* 'every', *kto-nibud* 'anybody' tend to be used within unique mentions in the discourse. The most reliable feature for different types of pronouns is the negative pronoun type. This result goes in hand with theoretical assumptions that free-choice pronouns

and negative pronouns are non-referential expressions and, thus, they are less probable in in a coreferential chain.

4 Discussion

The first experiment on the singleton detection has shown that the basic set of features gives relatively high precision and a low recall. Additional features (those taken from previous research for English as well as, relevant for first-mention detection) improve the recall, though there is a small precision loss. As a result, at least 70% of singletons can be filtered out from the mentions set for coreference pairs generation in coreference resolution task which will drastically improve computational efficiency and should improve the resolution quality.

5 Conclusions

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

The focus was on the testing features introduced for the corresponding task for English in the previous work, as well as some features used for the discourse-new detection.

We tested classifiers that can distinguish singletons. We also set a baseline for further experiments and tested special lexical features. Though the recall is low we think that the results can be used in the coreference resolution systems to filter out false candidates for the coreferring mentions.

The analysis of the results has shown that some of the lexical features such as special class of pronoun types, for example, negative and free choice pronouns are quite promising for this task and need further investigation and enhancing. The other promising direction is a more detailed investigation of the impact of the grammatical role of a given NP. In our future work we are planning to examine the contribution of more elaborated features.

Acknowledgments

The authors would like to thank the Lomonosov Moscow University students who participated in the corpus markup, Dmitrij Gorshkov for creating and maintaining corpus annotation software, and Dmitrij Privoznov for his useful comments.

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

Список литературы

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. Batista, G.E., Bazzan, A.L., Monard, M.C.: Balancing training data for automated annotation of keywords: a case study. In: WOB. pp. 10–18 (2003)
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. 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)
6. Hirschman, L., Chinchor, N.: Appendix f: Muc-7 coreference task definition (version 3.0). In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998 (1998), http://www.aclweb.org/anthology/M98-1029
7. 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)
8. Mitkov, R.: Anaphora resolution: the state of the art (1999)
9. 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)
10. 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)
11. Poesio, M., Ponzetto, S.P., Versley, Y.: Computational models of anaphora resolution: A survey (2010)
12. Prince, E.F.: The zpg letter: Subjects, definiteness, and information-status. Discourse description: diverse analyses of a fund raising text pp. 295–325 (1992)
13. 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)
14. 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)
15. 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)
16. Toldova, S., Ionov, M.: Features for discourse-new referent detection in russian. In: Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016, Konya, Turkey, April 3-9, 2016, Proceedings. Konya, Turkey (in press)
17. 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)

18. Uryupina, O.: High-precision identification of discourse new and unique noun phrases. In: ACL Student Workshop. Sapporo (2003)
19. Ward, G., Birner, B.: Information structure and non-canonical syntax. The handbook of pragmatics pp. 153–174 (2004)