The Third Personal Pronoun Anaphora Resolution in Texts from Narrow Subject Domains with Grammatical Errors and Mistypings

Daniel Skatov and Sergey Liverko

Dictum Ltd, Nizhny Novgorod, Russia
{ds,liverko}@dictum.ru

Abstract. The third personal pronoun anaphora resolution in texts from the Internet sources (forum comments, opinions) with a given subject domain (cars, household appliances etc) is being discussed. A concrete solution to the task is offered. High precision with acceptable recall (and vice versa) is shown by an example of opinions about mobile phones.

Keywords. Computational linguistics, natural language processing, anaphora resolution, machine learning, opinion mining.

1 Introduction

The problem of the third personal pronoun anaphora resolution discussed in this paper consists in the replacement of pronouns such as *"he", "his", "her", "it", ...* with nouns (antecedents) that these pronouns were used instead. Its solution is needed firstly in text mining applications, such as opinion mining (about goods, people) or fact extraction. Without resolved anaphoras those applications lose in recall of their results. The loss degree depends on the type of proceeded texts: e.g., in opinions about goods the density of *"it"* (masculine gender in Russian) pronoun is 1,5 times higher than in news[1].

The known methods of anaphora resolution can be divided into two groups — (1) statistical and (2) syntactical. Methods from class (1) [3] are based on the results of machine learning and are potentially applicable to texts of significantly different nature. Class (2) [1,2] exploits the sentence syntactical parsing tree (or semantic graphs as their derivatives) and as a result the applicability of such methods is limited to relatively «correct» texts (e.g., dossier texts [2]). This article describes a method combining these two approaches in a certain sense.

[1] A random sample of news from [12] (the anaphora density — 0.34 per 1 K and a sample of opinions about mobile phones from the sources such as [13] (the anaphora density — 0,53 per 1 K were used to perform measurements, each one of 1 Mb.

Texts from «real life» are full of typos and specialized slang with their grammar far from correct one:

Ive got a **whit ceise** and buttons peel **gradauly** and they becomes gray no cleaning helps or anything **likethat..!** Weak processor also made upset as well as small memory amount, it works terribly slow. (1)

The method of anaphora resolution, offered by the authors, takes mistypings and the results of syntactic parsing of text fragments (with mistypings corrected) into account. It is adapted to process texts from specific subject domains. Method can work with «correct» texts as well as informal ones (such as opinions or notes). To achieve a high processing quality for texts from a selected domain, a preliminary adjustment to the method is needed. It consists in learning on an unmarked corpus and composing the operating terminological dictionaries.

Three modes of the method have been implemented:

(A) good precision (70-80%) with high recall (90-95%),

(B) approximately equally good precision and recall (75-85%),

(C) excellent precision (up to 95%) with high acceptable recall (40-50%).

The implementation of the technology is represented by a software module called DictaScope Anaphora. It is adjusted to processing opinions about mobile phones from Internet sources. Within the bounds of the article an estimation of recall-precision ratio for processing such kind of data is carried out. The model is being used in the real application for online opinion monitoring. Modes A, B and C were obtained in the process of looking for a solution effective for this application – i.e. the one with high precision on possibly intentionally reduced input data.

2 Problem statement

Basic statement. For each pronoun pr_i, $i = 1,\ldots,N$ from text T choose the resolving pronoun (antecedent) a_i. *Remark.* In certain cases it is impossible to choose a_i, e.g.:

This mobile phone has a sensor screen. It's very inconvenient. **(screen or phone?)** (2)

Resolving of such an ambiguity (which can conditionally be called semantic) is a hard task even for a human, as both variants are of equal possibility. In the current problem statement it is offered either to choose a concrete antecedent or not to resolve the anaphora.

Advanced statement. It sometimes turns out that an acceptable precision of selecting a sole variant is unreachable. Therefore the following task specification is proposed: for each pronoun pr_i, $i = 1,\ldots,N$ form a list of possible resolving variants $\left(a_i^1,\ldots,a_i^{l_i}\right)$ sorted in accordance with their ranks (the first one is the best). Then a_i^1

can be chosen as a_i. In case a requirement of a high recall takes place (e.g., for posterior hand processing of results) it is sufficient to ensure high quality of ranking.

The variants of resolving antecedents can be supplied with real-value weights $w = w\left(a_i^k\right) \in (0,1]$, $i \in \{1,\ldots,N\}$, $k \in \{1,\ldots,l_i\}$, which correspond to each variant's confidence.

Traits. Let's resort to an example to make the task statement clear:

```
bought it for business, very useful because [it] {* =
0.652166, business = 0.2371, NULL = 0.168611} supports
two sim cards. Nice, big display, no dead spaces found
on [it]{display = 0.466248, * = 0.284525, NULL =
0.0777368, business = 0.0101848} (3)
```

For pronoun $pr_1 = «it»$ the list of variants is formed ($a_1^1 = «*»$, $a_1^2 = «bussiness»$, $a_1^3 = «NULL»$) with weights $w\left(a_1^1\right) \approx 0.65$, $w\left(a_1^2\right) \approx 0.237$, $w\left(a_1^3\right) \approx 0.1686$ (similarly for $pr_2 = «it»$). There are also special «*» and «*NULL*» designations:

- «*» — «the current object of discourse», so-called «implicit» antecedent. This is typical for opinions and reviews — i.e. for texts representing direct speech in writing. In the example above the word *«phone»* (as well as its concrete model reference) is not found anywhere before $pr_1 = «it»$, though the teller means exactly *«this phone»*.
- «*NULL*» — a directive «not to resolve pronoun». If «*NULL*» is at first position in the list of variants, the pronoun is left unresolved.

Thus, there are two cases in a basic problem statement in which the anaphora will not be resolved:

1. No variants for pronoun resolution is found;
2. «*NULL*» is the first in the ranged list of variants. It is easy to see that if, in case of semantic ambiguity, the probability of the correct choice of antecedent is less than ½, the precision will not fall on the average. Therefore, in this case the choice of «*NULL*» variant is justified.

In the example (3) the task in the basic statement is resolved correctly by choosing the first variant for each pronoun. A solution in a basic statement will be further estimated.

3 Review

The subject area of this paper is covered in the works of three Russian groups.

1. Ermakov A.E., RCO. In [2] empirical regularities of persons referencing are shown for texts from Russian mass media; they can be used to build a mechanism for

anaphora resolution in text sources of this class (with the help of natural language syntactic parser).

2. Tolpegin P., Vetrov D., Kropotov D. Article [3] describes an experience of this group in resolving the third personal pronoun anaphora in news by machine learning methods. The approach is typical for this type of solvers, the precision shown equals 62% on a control collection.
3. Okatiev V., Erechinskaya T., Skatov D. In the report [1] it is shown how pronoun anaphoras of different types can be resolved with the help of syntax parsing trees analysis. This approach is well applicable to the texts in which most of the sentences allow building correct syntax trees.

The specificity of this article — processing texts from narrow subject domains with mistypings and slang — is not touched upon in the works listed above.

The question discussed is more widely represented in foreign scientific works:

- from English-speaking authors patented system [11] and work [8] (which demonstrates values of basic indicators at a level about 80% while using probability model) are first to be mentioned;
- authors of [9] use maximum entropy method to resolve the third personal pronoun anaphora in Chinese, with F-measure about 70%;
- [10] describes an application of machine learning to personal pronouns anaphora resolution in Turkish with recall-precision at about 60-70%.

The overall impression of these works is the following: competent combination of analysis methods and rather full vocabulary data results in recall-precision not less that 70%.

4 Solution

4.1 Lists of variants and attributes

After tokenization (when the lists of grammar values of the tokens are supplemented taking mistypings into consideration) and dividing text into "conditional" sentences all the pronouns are looked through in the text from left to right. A concrete pronoun pr is fixed, $i = 1,\ldots,N$, and list $\mathrm{var}(pr)$ of possible antecedents is formed:

1. from all the words located within $\kappa = 2$ sentences to the left of pr, nouns in concordance with pr_i by gender and number are selected;
2. from the same words pronouns which are in concordance with pr by gender and number are selected and the list $\mathrm{var}(pr)$ is supplemented with nouns that resolve these pronouns.

Possible antecedents can also be found **to the right** of pr; however, not more than 30 examples of this were found in the corpus, with the correct variant also found to

the left of *pr* in ⅓ cases. Therefore, the possible variant location to the right is ignored by the method.

The proposed scheme has a chain character: pronouns on the left of given pr, which are close to it and already resolved, add antecedents which are located to the left of the boundary of the window $\mu = 2$ to $\mathrm{var}(pr)$. The scheme presents a certain compromise: the list can be imprecise but $\mathrm{var}(pr)$ remains quite compact. Advancing the window border κ up to 5 with the chain scheme disabled has led to a noticeable decrease in the solution precision during the experiments, so the decision was made to reject the varying left border.

For the further ranking of the lists $\mathrm{var}(pr)$ a vector of attributes $A(a)$ is calculated for each $a \in \mathrm{var}(pr)$. Let us mention the following attributes from the operational ones:

- $IsVoc \in \{0,1\}$ — the belonging of a to a terminological dictionary *TermVoc*
- $Freq \in \mathbb{N} \cup \{0\}$ — the number of mentionings of the given word (in any form) to the left of *pr* ;
- $Dist \in \mathbb{N}$ — the distance between the pronoun pr and the position of a inside the text (measured in words);
- $IsVerb \in \{0,1\}$ — the presence of direct father in a form of verb in syntax tree for a fragment containing a ;
- $NumNodes \in \mathbb{N} \cup \{0\}$ — the number of nodes in a bush subordinate to a .

The last two attributes have been introduced based on exploring correlation between numeric properties of a tree and resolving antecedents. For example, greater *NumNodes* were often correspondent to proper variants of resolution. These attributes values are set into null in case the tree was not formed.

The distance is measured in words for a number of reasons: (a) to get a valid syntactical unit (clause, noun phrase) was not possible (at that moment) due to the laboriousness of the adaptation of the syntactical parser to the special features of input texts (e.g. the absence of punctuation); (b) a paragraph is too large for being a unit of measure — the majority of opinions consist of one paragraph; (c) windows are measured in sentences and a two-sentence diapason is considered to be sufficient for the research.

IsVoc attribute implements the following idea: taking a subject domain's specificity into account allows to obtain higher quality of analysis. In fact, *IsVoc* allows to raise the priority of variants relating to subject domain of the text — they are of most interest (not always, though).

4.2 The test corpus

To evaluate the work of the methods a corpus of 3M was built from opinions about mobile phones from the sources like [13,14,15]. Due to the specificity of the application the corpus was additionally divided into three groups: positive, negative and neutral opinions, each of 0.8–1.2 M. As a next step it was marked up with the resolved anaphoras according to the following scheme:

- if the correct antecedent could be chosen directly from the text, its occurrence which was closest to the left of the pronoun being resolved was marked in a special way;
- in case of semantic ambiguity the pronoun was marked with «*NULL*» variant;
- the resolving word was written next to the pronoun in the corresponding case.

The statistical characteristics of the corpus were estimated.

- The whole number of 8.3 thousand opinions formed of 37 thousand unique word forms (including mistypings).
- The most frequent opinion length varying from 15 to 35 words; average opinion length — 54 words; the bulk of the opinions containing 10 to 90 words; opinions of more than 100 words are rare. The length scatter — from 2 to 340 words (Fig.1).
- Opinions consisting of one sentence are the most frequent; average opinion length — 4 sentences. The majority of opinions include 1 to 16 sentences; lengths more than 24 sentences are very rare (Fig.2).
- The corpus contains about 6.2 thousand third personal pronouns, including 4.5 thousand ones of masculine gender, 0.8 thousand of feminine gender, 0.7 thousand of plurals. The reason for a great number of masculine pronouns is the subject of the opinions (mobile phones).
- Less than 50% of the opinions do not contain any of the pronouns under research. 35% contain only one pronoun, about 10% — two of them. The maximum is 9 pronouns per opinion (Fig.3).

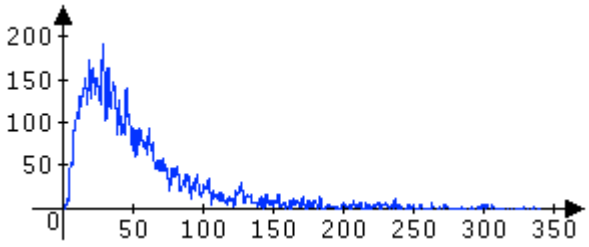

Fig. 1. Distribution of opinions lengths in words

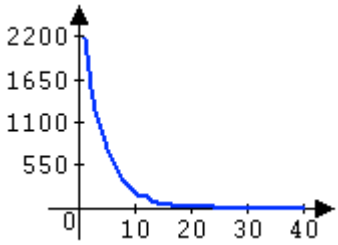

Fig. 2. Distribution of opinion lengths in sentences

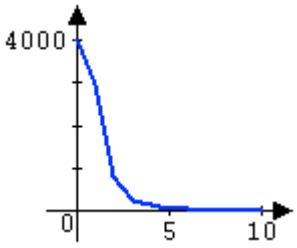

Fig. 3. Opinions distribution by a number of pronouns

4.3 Lexicographical analysis method

At the initial stage of studying a heuristic method for the options ranking was implemented:

- a system of priorities is formed on the set of attributes, which were listed in subparagraph 4.1;
- attribute values for each option are sorted according to the priorities;
- options are sorted lexicographically according to their sets of attributes.

The method resolves all the anaphoras for which it has found variants to the left with precision rate not more than 60%. The experiments in introducing new attributes and varying their priorities were not efficient. This has led the authors to the idea of filtration of the input data in order to achieve higher precision rate.

4.4 SVM-method based on machine learning

Let there be a general set of objects Y, divided into previously unknown classes, and a sample set $O \subset \Omega$, for each element of which its class is known. The task of classification is to answer the question: *"which class does each object* υ *from* Y *belong to"*, knowing only the sample set O (or the probabilities of belonging).

Let us fix a list $\mathrm{var}(pr_i)$ for one specific pronoun pr_i. In this case $O_i = \{A(a) \mid a \in \mathrm{var}(pr_i)\}$, $i = 1,\ldots,N$, and two classes are of interest — "are antecedents" and the inverse to it. Then the first class distance can be taken as $w(a)$.

Now we need to generalize the approach for N pronouns. Each set O_i represents an independent group, each of which consists of two classes — *"is the antecedent for pr_i"* and the inverse one, $2N$ classes for the whole training set. It is impossible to use this classification in practice with a different number $Q \neq N$ of *other* pronouns. In order to get exactly two classes for any number of pronouns, it is necessary to construct an acceptable combination of these groups. For this purpose, the authors propose adding attributes characterizing the group to each set $\omega_i \in O_i$. Thus within the same group all its members are additionally provided with the same set of numbers describing the group. The centroid can be taken as these numbers.

After expanding of the group members a sample set $\overline{O} = \bigcup_{i=1}^{N} O_i$ with the corresponding universe $\overline{\mathrm{Y}}$ and a fuzzy classifier $K(\omega) \in (0,1]$ which determines a distance between υ and the class *"are antecedents"* are constructed.

$K(\upsilon)$ is constructed in a form of so-called probabilistic decision function as described in [5,6] based on a classical C-SVM with a nonlinear kernel [7]. Selection of the core and the constants for the SVM was performed by minimizing the overtraining on the parameters grid while verifying the recall-precision ratio on the training and control samples. In the end, the kernel was chosen to be a polynomial one with a small degree.

Centroids raised the precision of the SVM-method from 70% to 80% (mode A).

4.5 Recall-precision regulator

To reach the precision rate of 90% linear discriminative analysis [4] was used: its aim is to find a line between classes, in the projection on which they are most discernible. With the help of discriminant, pronouns which may be not resolved (for the purpose of rising the precision rate) were identified. The combination of this filtration and SVM-method allowed to reach the desired result (mode C). Along the way, it was managed to derive mode B in which basic rates are balanced in the region of 75-85%.

5 Analysis of the results

5.1 Quality requirements and evaluation

Processing of the input set containing L third personal pronoun anaphoras is carried out in 2 steps.

1. **Filtration of anaphoras.** From the total number of L objects those for which the algorithm: (1) failed to form the set of variants, (2) put «NULL» in the first place in the list of variants or (3) eliminated from the examination due to regulator work are deleted. As a result, N anaphoras are left, for each of them the algorithm can choose an antecedent (not necessarily the correct one). If the whole of L anaphoras resolved *correctly* are considered as relevant, the recall rate of this step is N/L while the precision is equal to 1, as all chosen objects (N) are included in the relevant (L).
2. **Resolution of the left anaphoras.** In this step the whole of N anaphoras resolved correctly are considered as relevant. The algorithm attempts to resolve them, succeeding in K cases. Due to the coincidence between the volumes of relevant objects and those being resolved, the precision and recall rates are both equal to K/N.

Two out of four rates mentioned above (precision and recall for each step) are informative:

- recall is a portion of pronouns for which the algorithm succeeded in finding an antecedent;
- precision is equal to a percent of this portion containing correctly identified antecedents.

To the writers' opinion, this approach to evaluation conforms to the quality requirements. In addition, the estimations do not depend on the mechanism of anaphora resolution (including the size of variant lists).

5.2 The quality of SVM-method and sensitivity to the sample volume

Opinions containing at least one of the pronouns under research (4 thousand altogether) were selected from the corpus. To evaluate the SVM-method sensitivity to the sample volume this set of opinions underwent the procedure of q-fold cross validation.

Verification was carried out for $q = 1,\ldots,300$, i.e. $q = 1$ means verification of the model for the whole 4 thousand opinions, $q = 300$ — for a sample of 13 opinions. For each q the mean of recall and precision was calculated for each iteration as well as their minimum and maximum for the diagrams reflecting the dependency between quality and the volume of input data.

Measuring was done for modes A, B and C (Fig.4, abscissa corresponds to q).

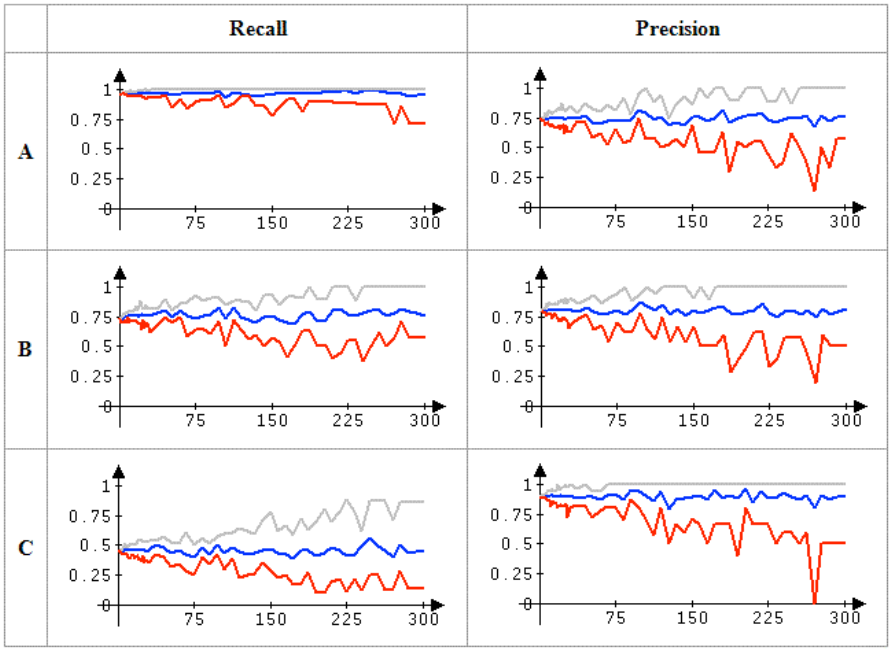

Fig. 4. Results for SVM-method cross-validation in A,B,C modes

It can be seen that all the means are stable even for small-sized samples.

Table 1. Averaged quality measures for SVM-method

	Recall	**Precision**
A	97.3%	74.2%
B	75.4%	80.7%
C	45.6%	90.3%

5.3 The results of ROC-analysis of SVM-method

Fig.5 illustrates ROC-curves for SVM-method in A, B and C modes.

The area under **A** curve is 0.74, under **B** one — 0.76, which is considered as "good" according to the expert scale. The area under **C** curve is 0.81 with this mode considered as "very good".

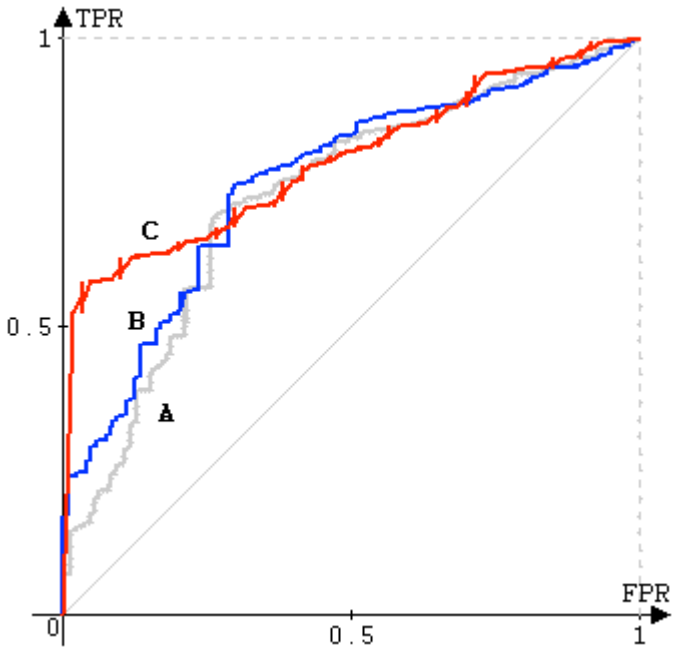

Fig. 5. ROC-curves for SVM-method in A, B, C modes

5.4 The SVM-method independence of the sentiment of the corpus

It was additionally verified in empirical way that the SVM-method is independent of the sentiment of the texts processed, since it cannot be forgotten that anaphoras in negative opinions might be different from those in positive opinions.

The "negative" corpus was used as a training set, the "positive" one as a control set.

Table 2. Check for SVM-method independency from sentiment

(RECALL %, PRECISION %)	**(A)**	**(B)**	**(C)**
Negative *(training)*	(95.1, 80.2)	(77.8, 86.7)	(43.1, 93.2)
Positive *(control)*	(96.3, 78.7)	(79.1, 83.4)	(56.2, 89.9)

5.5 Significance of the factors

Discriminative analysis provides an estimation of contribution of the attributes to the common decision — the judgment can be made based on the coefficients for the corresponding attributes in the linear discriminant and the range of attribute values. It

is also possible to estimate how much influence components of the centroid bring to the solution.

According to the Table 3, the frequency is two times more important than the distance, the presence of a father-verb is more important than the number of nodes in the bush (even if correcting this by a wide range of *NumNodes* — sometimes up to 10-15 knots). Picture according to the centroid is consistent on a whole, except for *IsImp* and *IsVoc*, so their contribution can be estimated to be approximately equal.

Table 3. Valuing the attributes significance according to the results of discriminant analysis

Attribute	*Coefficient in linear discriminant*	*Corresponding coefficient near the component of the centroid*
IsImp $\in \{0,1\}$	– 2.9	18.8
IsVoc $\in \{0,1\}$	9.3	1.1
HasVerb $\in \{0,1\}$	– 7	35.8
NumNodes $\in \mathbb{N} \cup \{0\}$	– 0.5	18.9
Freq $\in \mathbb{N} \cup \{0\}$	– 21.5	-1.6
Dist $\in \mathbb{N}$	– 10.6	0.1

Compiling vocabularies for *IsVoc* is rather laborious. The authors have discovered that the main coefficients in modes A and C (recall and precision respectively) reduce from about 90 to 70% when this attribute is not used; in mode B both coefficients reduce by ~10%. It can be stated that it is precisely *IsVoc* attribute that allows to achieve the precision rate of 90% and higher.

5.6 Evaluation of lexicographical method

The advantage of this method is that no marked-up corpus is needed for its initialization. The practical use of the SVM-method has shown that a trained classifier copes with texts from domains different from that of the training set with the rates declining by several percents (with the exception of *IsVoc* attribute — new vocabularies are needed).

Table 4. Estimation of the lexicographical method quality

	With IsVoc	**Without IsVoc**
(RECALL %, PRECISION %)	(93.7, 51.9)	(93.7, 42.4)

The main error f the method is an excessively strong influence of an attribute with the highest priority. E.g. using *IsVoc* attribute often results in an incorrect choosing a vocabulary word while not using it — in choosing the word closest to the left.

6 Conclusion

This paper offers a solution to the problem of the third personal pronoun anaphora resolution. The software complex called DictaScope Anaphora was implemented based on the models and methods discussed in this paper. It has the following characteristics:

- there are three modes, which allow to achieve both recall and precision rates of 80% or to give preference to one of them and achieve the result of 95%;
- it is possible to take mistypings and grammatical errors into account, which is important for processing texts from online sources (such as reviews);
- in this case an adjustment of the parameters for a specific subject area is needed.

The features of the internal structure of the system and the mathematical foundation are described; the detailed evaluation of the test data and the quality of its processing is carried out.

Among the shortcomings it is a drop in accuracy on the masculine pronouns that should be noted. It is caused by the choice of the subject of opinions (a mobile phone). It is mentioned very often (including implicit mentioning) and the main part of malfunctions consists in choosing an implicit antecedent «*». In authors' opinion, the problem can be solved by taking new attributes connected with the result of syntactical parsing into consideration.

The development plans include the application of the system to other domains and improving the recall-precision ratio by introducing new attributes and refining the adjustment of the coefficients.

7 References

1. Okatev V.V., Gergel V.P., Alexeev V.E., Talanov V.A., Barkalov K.A., Skatov D.S., Erekhinskaya T.N., Kotov A.E., Titova A.S. Report on research implementation on the topic: "Development of a pilot version of syntactical analyzer for the Russian Language", VNTIC Inventory Number 02200803750 // VNTIC, Moscow (2008)
2. Ermakov A.E. Referencing the designations of persons and organizations in Russian media texts: empirical laws for computer analysis. In: Proceedings of the International Conference "Dialog'2005", Computational Linguistics and Intelligent Technologies (2005)
3. Tolpegin P.V., Wind D.P., Kropotov D.A. Algorithm for automated third-person pronouns resolution on the basis of machine learning methods. In: Proceedings of International Conference "Dialog'2006", pp. 504-507. Izd RGGU, Moscow (2006)
4. Oldenderfer M.S., Blashfield R.K. Factor, discriminant and cluster analysis. Under. Ed. Igor Enyukova. Finance and Statistics, Moscow (1989)

5. Platt John C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In Advances in Large Margin Classifiers, Alexander J. Smola, Peter Bartlett, Bernhard Sch olkopf, Dale Schuurmans, eds., MIT Press, (1999)
6. Hsuan-Tien Lin, Chih-Jen Lin, Ruby C. Weng, A note on Platt's probabilistic outputs for support vector machines. In: Machine Learning, v.68 n.3, p.267-276 (October 2007)
7. Vapnik V. Statistical Learning Theory. Wiley (1998)
8. Niyu G., Hale J., Charniak E. A statistical approach to anaphora resolution // In; Proceedings of the Sixth Workshop on Very Large Corpora. COLING-ACL'98. Montreal, Canada (1998)
9. Ning Pang, Jun-feng Shi. The third personal pronoun anaphora resolution in the paroxysmal text of the Chinese web. In. Coll. of Appl. Sci., Taiyuan Sci. & Technol. Univ., Taiyuan, China
10. Yıldırım S., Kılıçaslan Y. A machine learning approach to personal pronoun resolution in Turkish. In Proceedings of 20th International FLAIRS Conference, FLAIRS-20. Key West, Florida (2007)
11. Michael P., Kazuhide Y., Eiichiro S. Anaphora analyzing apparatus provided with antecedent candidate rejecting means using candidate rejecting decision tree. Patent US6343266 (2002)
12. Novoteka — news of the day: http://www.novoteka.ru.
13. Yandex.Market — search, selection and purchase of goods: http://market.yandex.ru.
14. CNews Internet-portal: http://zoom.cnews.ru.
15. All for Nokia phones: http://www.allnokia.ru.