Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2017"

Moscow, May 31—June 3, 2017

# RHETORICAL STRUCTURE THEORY AS A FEATURE FOR DECEPTION DETECTION IN NEWS REPORTS IN THE RUSSIAN LANGUAGE

**Pisarevskaya D.** (dinabpr@gmail.com)

Institute for System Programming of the RAS, Moscow, Russia

The framework of the Rhetorical Structure Theory (RST) can be used to reveal the differences between structures of truthful and deceptive (fake) news. This approach was already used for English. In this paper it is applied to Russian. Corpus consists of 134 truthful and deceptive news stories in Russian. Texts annotations contain 33 relation categories. Three data sets of experimental data were created: with only rhetorical relation categories (frequencies), with rhetorical relation categories and bigrams of categories, with rhetorical relation categories and trigrams of categories. Support Vector Machines and Random Forest Classifier were used for text classification. The best results we got by using Support Vector Machines with linear kernel for the first data set (0.65). The model could be used as a preliminary filter for fake news detection.

**Key words:** deception detection, Rhetorical Structure Theory, automated decepted detection, news verification, discourse analysis

## 1. Introduction

In the contemporary world we deal with the large amount of information that we get from different and diverse sources: newspapers, institutional and non-institutional online media, blogs and social media, TV channels and their websites etc. It is very important to understand the difference between information types and to evaluate reliability of sources. In news reports, rumours, deceptive information and deceptive (fake) news can be easily used for manipulation of public opinion, for information warfare. This is why new tools for automated deception detection and information verification, created for different languages, based on Natural Language Processing methods and models, are required in our society. Now there are no research

papers about automated deception detection for the Russian language. There is also a significant lack of linguistics tools for Natural Language Processing which could be helpful in solving the problem. It seems to be a theoretical and methodological challenge.

## 2. Literature Review

Written texts are a subject of research for studying deception detection methods, especially for English. Digital texts, online reviews (Ott et al., 2011; Mukherjee et al., 2013), fake social network profiles (Kumar, Reddy, 2012), fake dating profiles (Toma, Hancock, 2012) etc. were already investigated. The objective of revealing news verification mechanisms arose rather recently. Fake news may be identified on different levels. Usually researchers tend to combine different levels, from lexics and semantics to syntax. Most studies focus on lexics and semantics and some syntax principles; discourse and pragmatics have still rarely been considered (Rubin et al., 2015) due to the complexity of such approach.

On the lexics level researchers can extract some stylistic features (part of speech, length of words, subjectivity terms etc.) that help to apart tabloid news (they are similar to fake news) with 0.77 accuracy score (Lex et al., 2010). Numbers, imperatives, names of media persons can be extracted from news headlines (Clark, 2014); the numbers of these keywords can be used as features for classification with Support Vector Machines or Naive Bayes Classifier (Lary et al., 2010). Some linguistics markers can be found in lexics and semantics level from the Statement Validity Analysis, the accuracy rate reaches 0.74 (Porter and Juille, 1996). Existing psycholinguistics lexicons, for instance LIWC (Pennebaker and Francis, 1999), can be used in performing binary text classifications for truthful vs deceptive texts (0.70 accuracy rate) (Mihalcea and Strapparava, 1999)—for example, methods can be based on frequency of affective words or action words from lexicons. As to syntax level, Probability Context Free Grammars can be used. Text fragments are presented as a set of rewrite rules to describe syntax structure and produce a parse tree. So we can distinguish rule categories for deception detection with 0.85–0.91 accuracy (Feng et al., 2012). Syntax analysis is often combined with other linguistics or network approaches (Rubin et al., 2015a). On pragmatics level, it is found out that pronouns with antecedents in text are more often used in fake news‘ headlines to pay reader’s attention (Blom and Hansen, 2015).

Some studies are focused on creating models that reveal if the described event accords with the facts or not. In (Sauri and Pustejovsky, 2012) authors represent a model, which is based on grammatical fact description structures in English and kindred languages. It has been implemented in De Facto, a factuality profiler for eventualities mentioned in text based on lexical types and syntax constructions. The researchers also created FactBank—annotated corpus in English.

There are three types of fake news: serious fabrications, large-scale hoaxes and humorous fakes (Rubin et al., 2015b). We should also take into consideration the recent research (Hardalov et al., 2016): it proposes the approach for automatically distinguishing credible from fake news, based on different features: linguistic (n-gram), credibility-related (capitalization, punctuation, pronoun use, sentiment

polarity), and semantic (embeddings and DBPedia data) features. The accuracy is from 0.75 to 0.99 on three different datasets, but, although the approach, based on combining different levels, is promising, it is hard to compare the results because they are got mainly on news stories which may be considered as hoaxes and fakes (7038 fake texts), not as intentiallly fabricated "serious" news (68 fake texts).

Recent research projects are dedicated to discourse differences between deceptive (fabricated) and truthful (authentic) news, specifically in terms of their rhetorical structures and coherence relation patterns (Rubin et al., 2015). Vector space modeling application lets predict whether a particular news report is truthful or deceptive (0.63 accuracy) for English. Seriously fabricated news stories were chosen for the dataset. So rhetorical structures and discourse constituent parts and their coherence relations are already reviewed as possible deception detection markers in English news. If we review deception detection methods for other languages, in our case for Russian, we also should keep in mind linguistics and cultural considerations.

RST (Rhetorical Structure Theory) framework (Mann and Thompson, 1988) is addressed to the discourse level of text. It represents text as an hierarchical tree. Some parts are more essential (nucleus) than others (satellite). Elementary discourse units are connected to each other according to relations: elaboration, justify, contrast etc. The theory pretends to be universal for all languages, therefore we chose it for our research. It is used in Russian computational linguistics. Nevertheless automated parser was never worked out specially for the Russian language, it causes constraints in using RST framework in applications.

Support Vector Machines (SVM) method can be grasped as supervised learning models for classification tasks in machine learning. In our case, news reports are shown as vectors in n-dimensional space. A news report is placed in one of two groups, deceptive or truthful. Random forest is as a learning method operates by constructing a multitude of decision trees at training time and, in case of classification task, outputting the class.

## 3. Research Objective

Our hypothesis is that there are significant differences between structures of truthful news reports and deceptive ones. Our aim is to reveal them using RST relations as deception detection markers, based on the definite corpus. Firstly, we would like to find out what the features from the Rhetorical Structure Theopy should look like: we should detect if RST relation types' frequencies, relations' sequences are important. Then we shall estimate the impact of these features into the successful detection. We shall classify the texts, based on the RST relations labeling, and we shall do our best to predict if news reports are truthful or deceptive.

This model can be useful for news verification, in detecting and filtering deceptive (fake) news. Especially it is of vital necessity for the Russian language, because news reports in Russian nowadays often contain deceptive information and deliberate misinformation, and there is no way how to check it excepting the manual one. Our research is based on the methodology of the news reports research for the English language (Rubin et al., 2015), but it also takes into consideration some features of this research field for Russian.

## 4. Data Collection

The main difficulty of collecting data set for deception detection is the lack of sources in Russian that contain verified samples of fake and truthful news. There are no Factbanks in Russian, there are no objective, impersonal fact checking websites that contain the reports of investigative journalism. Therefore, the only way out in solving the problem was the reliance on the presented facts, on the factuality. The daily manual monitoring of news lasted 11 months (June 2015-April 2016). Online newspapers in Russian were used as sources. In order to maintain balance we took texts from different sources: well-known news agencies' websites, local or topic-based news portals, online newspapers from different countries (Russia, Ukraine, Armenia etc.). News source mention was not included in corpus text annotations to avoid subjectivity. Blog texts, social media content, news reports based on opinions (not on facts) were excluded from the monitoring. So we used news reports about facts and not analytic journalism stories where different viewpoints are conventional. News stories were analyzed in retrospect when the factuality was already known. Fake reports were put to negative class ('0'), truthful reports were put to positive class ('1').

For instance, news story about airplane crush which appeared only in one source and did not fit facts was considered as fake. News story about death of famous person was condidered as fake after refutation. Airplane accident which was mentioned in diverse sources and confirmed with facts was considered as true. Death of famous person which was confirmed in other news sources by this person's friends and relatives was considered as true. So we see two news "pairs" about definite topics. They can be not only about the same topic, but about the same event: for example, news story about the Shengen visa centres closing for Russian citizens was considered as fake because at the same time we could see the truthful news story about new rules of document executions and possible delays.

As to news reports with mutual contradictions, a report was added to fake cases if we could see the opposite news reports at the same time in different online media: with some unproven facts and with their refutation which was truthful. It means that if we saw a fake news story we considered the time when it appeared: if there were stories with refutation at the same time, we considered that it was an intended fake and not a journalist's mistake caused by lack of facts.

There are three types of fake news: serious fabrications, large-scale hoaxes, humorous fakes (Rubin et al., 2015b). We analyzed only the first two types, because we are interested in deceptive news that look similar to truthful news. We suggest: if only a report is intended as a fake one, its rhetorical structure differs from a truthful one. That's why we did not add reports, based on author's inaccuracy and not on author's intention, to our corpus.

Generally, the final data set consists of news reports dedicated to 38 different topics, with equal number of truthful and deceptive news stories to each topic, and not more than 12 news reports about the same topic. Each topic was analyzed carefully to define a fake part in the case and to avoid subjectivity and biased evaluation.

## 5. Corpus Details

The corpus contains 134 news reports, with average length 2700 symbols. Average number of rhetorical relations in text is 17.43. The whole number of rhetorical relations in corpus is 2340. Clauses were taken as elementary discourse units.

For comparison, the dataset in the paper describing the research on which we base our research (Rubin et al., 2015) includes 144 news reports. Corpus in the research about the impact of discourse markers in argument units classification (Eckle-Kohler et al., 2015) consists of 88 documents, predominantly news texts. So the corpus size is conventional for our goals for the initial research on the field of discourse analysis.

There are no discourse parsers for Russian, that‘s why tagging and validation were made manually. We used UAM CorpusTool for discourse-level annotation. We based the research on the ”classic“ set by Mann and Thompson and added to it the relational categories from extended sets. News reports usually have a definite template, thus, we used a relatively small number of different relational categories. We created relation types Evidence 1 (the source of information, the speaker, is mentioned precisely without hyperlink), Evidence 2 (the source is mentioned imprecisely: «Some experts/media say that...»), Evidence 3 (the source is mentioned precisely with hyperlink) and Evidence 4, the most rare one (the source is mentioned with hyperlink, but the information in the source text does not correspond to the information in the news report). They have the same structure in text, but we guessed that there could be a difference between truthful and deceptive news. Finally we had 33 relation types: ’Circumstance‘, ’Reason‘, ’Evidence1‘, ’Evidence2‘, ’Evidence3‘, ’Evidence4‘, ’Contrast‘, ’Restatement‘, ’Disjunction‘, ’Unconditional‘, ’Sequence‘, ’Motivation‘, ’Summary‘, ’Comparison‘, ’Non-Volitional Cause‘, ’Antithesis‘, ’Volitional Cause‘, ’Non-Volitional Result‘, ’Joint‘, ’Elaboration‘, ’Background‘, ’Solution‘, ’Evaluation‘, ’Interpretation‘, ’Concession‘, ’Means‘, ’Conjunction‘, ’Volitional Result‘, ’Justify‘, ’Condition‘, ’Exemplify‘, ’Otherwise‘, ’Purpose'.

## 6. Inter-annotator Consistency

We faced the following discrepancies during our tagging work: Background/ Sequence/Elaboration; Reason/Unvolitional Cause/Volitional Cause; Purpose/ Unvolitional Result/Volitional Result; Evaluation/Interpretation; Antithesis/Contrast; Elaboration/Justify/Restatement in quotations. We prepared guidelines in our tagging manual for these cases.

The assignment of RST relations is often criticized because it could be connected with the subjectivity of annotators' interpretation: the same text could be annotated in different ways. We tried to solve this problem by preparing a precise manual for tagging and by developing consensus-building procedures. News topics for coders were selected randomly, after that coder A analyzed 66 reports, coder B analyzed the remaining 68 reports. Truthful and deceptive news reports about the same event were annotated by the same person. Therefore, if there could be a variance in segmenting a text into clauses or in tagging a definite rhetorical relation type, similar parts and mutual quotations in truthful and deceptive texts would be annotated in the same way.

We selected Krippendorff's unitized alpha as a measure to apply because it suits if coders have different approaches to segmenting and labeling in definite text sequences. After the second step the agreement reached 0.75.

## 7. Data Analysis

The first experiment allows to define a baseline on the lexics level: we decided to choose frequency of lemmas from a sentiment lexicon as a feature for each text. We suggested that it could help identify truthful and deceptive news reports because positive and negative opinion words could be considered as affective words and could replace causation in deceptive texts. We used a list of 5000 sentiment words got from reviews devoted to various topics (Chetviorkin and Loukachevitch, 2012).

The second experiment was run on three different datasets. RST relation types frequencies and their collocations are represented as features. The first dataset (model A) is based on a statistics file which contains data about types of RST relations and their frequencies for each news report. In fact, we deal here with a 'bag of relation types', disregarding their order. As rhetorical structure is tree-like and not flat, we added count of bigrams and trigrams of RST types (based on class nltk.util.ngrams in NLTK 3.0 (Natural Language Toolkit) for Python) for each text in model A to create model B. Model C also contains model A, but in this case it is combined for each news report with count of occurences of top 20 bigrams of RST types and top 20 trigrams of RST types from the whole corpus (here we used module nltk.collocations, threshold not less than 3 occurencies for the whole corpus).

We selected two supervised learning methods for texts classification and machine learning: Support vector machines (SVMs) and Random Forest, both realized in scikit-learn library for Python. SVMs were used with linear kernel and with rbf kernel. In both experiments we used 10-fold cross-validation for estimator performance evaluation.

We also held an additional experiment: the corpus was annotated manually to compare machine learning results, which are based on RST-features, with human asessments. 25 participants, aged 20–35, who did not participate in choosing texts for the coprus or annotating RST relations, marked per e-mail each news report as truthful/fake one (every participant marked all texts). We did not use online forms, because these people also gave expert interviews during preliminary qualitative sociological research about fake news perception, and it was convenient to discuss all issues per e-mail. After that we counted common scores.

## 8. Statistical Procedures

The results for the first and second experiments are presented in Table 1. We can evaluate that the classification task is solved better by SVMs (linear kernel) for model A, without addition of bigrams and trigrams features. The accuracy score is 0.65. It means that the sequence of RST relations is not so important as the frequencies of RST relation types. The score can be compared with the predictive power of the model for English (Rubin et al., 2015) which is 0.63. It is also more than the human

ability score to detect deceptive information (0.54) which was got in different experiments listed in the article (Rubin et al., 2015). The results for our additional experiment with manual tags are got together in Table 2. They can be compared with the results for English. They show less recall and less precision than the results of automated deception detection for Russian in our case.

The most significant features which influence on linear SVMs classification for model A are: 'Justify', 'Evidence3', 'Contrast', 'Evidence1', 'Volitional Cause', 'Comparison'. So we decided correctly to divide 'Evidence' into 4 types. Student's t-test to check the statistical significance of these six features showed that first five ones are significant, about 'Comparison' we cannot state the same with confidence (p-value measure 0.07858).

'Volitional Cause' is one of the most significant features, and this relation type is more typical for deceptive texts. Probably this could be explained so: authors of fake news pay more attention to the causation, because they want to explain an event with the internal logic of their position, without any inconsistencies. 'Circumstance' and 'Elaboration' are also more typical for deceptive news reports, and they also point to the logical structure of a text. Herewith, 'Volitional Cause' is not the most significant feature. 'Justify', 'Evidence3' and 'Evidence1', 'Contrast' are more typical for thuthful texts. Hence, truthful news reports contain more often information with rational, precise source mention and direct link to it (whereas 'Evidence2' is more typical for fake news, as it contains imprecise source mention). The presence of 'Contrast' and 'Comparison' among important features can be explained so: truthful news reports in our corpus can be considered as rebuttals of fake news reports, therefore they refer to them and contain parts of deceptive texts. 'Contrast' and 'Comparison' could be used as a link between a deceptive text citation and an explanation why it is a fake.

**Table 1.** Results for different classifiers

| | **Precision** | **Accuracy** | **Recall** | **F-measure** |
|---|---|---|---|---|
| **Support Vector Machines, rbf kernel, 10-fold cross-validation** | | | | |
| Baseline | 0.38 | 0.42 | 0.54 | 0.42 |
| Model A | 0.54 | 0.53 | 0.51 | 0.51 |
| Model B | 0.60 | 0.55 | 0.52 | 0.50 |
| Model C | 0.65 | 0.61 | 0.56 | 0.57 |
| **Support Vector Machines, linear kernel, 10-fold cross-validation** | | | | |
| Baseline | 0.23 | 0.37 | 0.49 | 0.31 |
| Model A | 0.64 | 0.65 | 0.65 | 0.63 |
| Model B | 0.64 | 0.60 | 0.48 | 0.53 |
| Model C | 0.62 | 0.59 | 0.60 | 0.59 |
| **Random Forest Classifier, 10-fold cross-validation** | | | | |
| Baseline | 0.48 | 0.48 | 0.55 | 0.49 |
| Model A | 0.56 | 0.54 | 0.45 | 0.47 |
| Model B | 0.60 | 0.63 | 0.56 | 0.56 |
| Model C | 0.57 | 0.55 | 0.46 | 0.49 |

**Table 2.** Manual (human) asessments for news reports

| | **Precision** | **Recall** | **F-measure** |
|---|---|---|---|
| Scores for human asessments | 0.55 | 0.46 | 0.50 |

## 9. Discussion

Automated deception detection based on the Rhetorical Structure Theory seems to be a promising and methodologically challenging research topic, and further measures should be taken to find features for deception/truth detection in automated news verification model for the Russian language. Our hypothesis is confirmed. The present research is initial, and the model should be developed and modified, learned and tested on larger data collections with different topics. In addition, we should use a complex approach and combine this method with other linguistics and statistical methods. For instance, syntactic level features on top of sequences of discourse relations should be studied. Discourse markers may be also taken into consideration as separate features. The guidelines for gathering a training corpus of obviously truthful/deceptive news should also be improved.

The extrapolation of the existing model to all possible news reports in Russian, devoted to different topics, would be incorrect. But despite this fact, it can already be used as a preliminary filter for deceptive (fake) news detection. Results of its work should be double-checked and refined, especially for suspicious instances fact checking.

We tried to take into consideration 'the trees'—hierarchies of RST relation types in texts and dependences between relation types. This aspect should be studied more deeply and intensively.

The model is also restricted by the absence of automated discourse parser for Russian. It is typical for other Natural Language Processing tasks for Russian which deal with RST.

Finally, the assignment of RST relations to news report could be connected with the subjectivity of annotators' interpretation. Despite of inter-annotator consistency measures, this problem exists and could be partly solved by preparing more precise manuals for tagging and by developing consensus-building procedures.

## 10. Conclusions

News verification tends to be a very important issue in our actual world, with its information warfare and propaganda methods. The precision of human deception detection ability for news reports in the present research in Russian is 0.55.

We collected a corpus (134 news reports, truthful and fake ones). We segmented the texts manually and applied RST relations tagging to them. As to the experiments, three dataset models for machine learning were based on features from the Rhetorical Structure Theory. We also used the model based on features from the sentiment lexicon as a baseline. We applied Support vector machines (SVMs) algorithm (linear kernel / rbf kernel) and Random Forest to classify the news reports into 2 classes: truthful/deceptive. The predictive power of the model based simply on frequencies

of RST relation types in texts is the highest one (the sequence of RST relations is not so important). The classification task is solved better by SVMs (linear kernel) for this dataset (0.65 accuracy score). Such RST relation types as Justify, Evidence3, Contrast, Evidence1, Volitional Cause, Comparison produce the most significant features. The modified model could combine RST relations markers with other deception detection markers in order to make a better predictive model.

## References


1. *Blom J. N., Hansen K. R.* (2015), Click bait: Forward-reference as lure in online news headlines, Journal of Pragmatics 76, pp. 87–100.
2. *Chetviorkin I. I., Loukachevitch N. V.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain, Proceedings of COLING 2012: Technical Papers, pp. 593–610.
3. *Clark R.* (2014), Top 8 Secrets of How to Write an Upworthy Headline, Poynter, URL: http://www.poynter.org/news/media-innovation/255886/top-8-secrets-of-how-to-write-an-upworthy-headline/
4. *Eckle-Kohler J., Kluge R., Gurevych I.* (2015), On the Role of Discourse Markers for Discriminating Claims and Premises in Argumentative Discourse, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2236–2242.
5. *Feng S., Banerjee R., Choi Y.* (2012), Syntactic Stylometry for Deception Detection, Proceedings 50th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Volume 2: Short Papers, pp. 171–175.
6. *Hardalov M., Koychev I., Nakov P.* (2016), In Search of Credible News, Artificial Intelligence: Methodology, Systems, and Applications, pp. 172–180.
7. *Kumar N., Reddy R. N.* (2012), Automatic Detection of Fake Profiles in Online Social Networks, BTech Thesis.
8. *Lary D. J., Nikitkov A., Stone D.* (2010), Which Machine-Learning Models Best Predict Online Auction Seller Deception Risk?, American Accounting Association AAA Strategic and Emerging Technologies.
9. *Lex E., Juffinger A., Granitzer M.* (2010), Objectivity classification in online media, Proceedings of the 21st ACM conference on Hypertext and hypermedia, pp. 293–294.
10. *Mann W. C., Thompson S. A.* (1088), Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, Text, vol. 8, no.3, pp. 243–281.
11. *Mihalcea R., Strapparava C.* (1999), The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language, Proceedings 47th Annual Meeting of the Association for Computational Linguistics, Singapore, pp. 309–312.
12. *Mukherjee A. et al.* (2013), Fake Review Detection: Classification & Analysis of Real & Pseudo Reviews. Technical Report, Department of Computer Science, University of Illinois at Chicago, & Google Inc.
13. *Ott M., Choi Y., Cardie C., Hancock J. T.* (2011), Finding deceptive opinion spam by any stretch of the imagination. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, pp. 309–319.

14. *Pennebaker J., Francis M.* (1999), Linguistic inquiry and word count: LIWC, Erlbaum Publishers.
15. *Porter S., Juille J. C.* (1996), The language of deceit: An investigation of the verbal clues to deception in the interrogation context, Law and Human Behavior, vol. 20, № 4, pp. 443–458.
16. *Rubin V. L., Conroy N. J., Chen Y. C.* (2015), Towards News Verification: Deception Detection Methods for News Discourse, Proceedings of the Hawaii International Conference on System Sciences (HICSS48) Symposium on Rapid Screening Technologies, Deception Detection and Credibility Assessment Symposium, January 5–8, Grand Hyatt, Kauai, 11 pages.
17. *Rubin V. L., Conroy N. J., Chen Y.* (2015a), Automatic Deception Detection: Methods for Finding Fake News, Conference: ASIS T2015, At St. Louis, MO, USA.
18. *Rubin V. L., Conroy N. J., Chen Y.* (2015b), Deception Detection for News: Three Types of Fakes. Conference: ASIS T2015, At St. Louis, MO, USA.
19. *Sauri R., Pustejovsky J.* (2012), Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text, Computational Linguistics, pp. 1–39.
20. *Toma C. L., Hancock J. T.* (2012), What Lies Beneath: The Linguistic Traces of Deception in Online Dating Profiles, Journal of Communication, vol. 62, № 1, pp. 78–97.