IJCLA VOL. 4, NO. 2, JUL-DEC 2013, PP. 189–202
RECEIVED 10/12/12 ACCEPTED 04/03/13 FINAL 13/03/13

Automatic Distinction between Natural and Automatically Generated Texts Using Morphological and Syntactic Information

LEONID CINMAN, PAVEL DYACHENKO, VADIM PETROCHENKOV, AND SVETLANA TIMOSHENKO

Institute for Information Transmission Problems, Russia

ABSTRACT

Our work lies in the field of automatic metrics for assessing text quality. However, the task we had to solve is different from the usual tasks of this domain. The traditional and most common formulation of the task is to distinguish well-written texts from poorly written ones, in which case it is presupposed that any text to be assessed is written by a human. Normally, the type of the text is also known: a scientific publication, news, etc. We set a more general task: to distinguish normal texts written by man, on one hand, from automatically generated texts or automatically processed and intentionally damaged natural texts, on the other hand. An additional difficulty is that "normal" texts in our collection contain lists, fragments of tables, and examples of bad texts with mistakes. We started by parsing our data with our syntactic parser for Russian, after which we trained an algorithm using words with extracted morphological and syntactic information. Our best results show 78.1% recall, 94.6% precision and 85.5% f-measure.

KEYWORDS: *Dependency parser, LibLinear, text quality, machine learning.*

1 Introduction

Our work lies in the field of automatic metrics for assessing text quality. Inside the domain we can see two streams of research – studies of readability and studies of coherence. The first one is presented, for example, in (Collins-Thompson, Callan, 2004), (Schwarm, Ostendorf, 2005). Papers by Barzilay, Lee (2004), and Soricut, Marcu (2006) can give an idea about the topics and methods in the second stream of studies. It is easy to see that while the researchers working on readability are focused on natural, human-written texts and their perception by other people, those who study text coherence work primarily with automatically generated texts. However, there are situations in which one has to process both automatically generated and human-written texts on the same principles: this will happen if the collection to be considered is heterogeneous.

To the best of our knowledge, there is only one recent paper dedicated to the uniform treatment of heterogeneous texts: (Louis, 2012). The author proposes to use genre-specific features to qualify texts, which means that at least we need to know beforehand what type of text we have – this is an indispensable condition for future treatment.

Our task, however, is different and simply formulated: we want to have an algorithm that could define whether a particular text is automatically generated (or automatically transformed from a natural text), or not. A simple question, but in a sense it may be considered as basic knowledge, which precedes any further processing.

An additional motivation for the experiment we are about to present is the situation in machine learning on Russian data. There is not much work done on Russian, besides, most of them report inferior performance for Russian than for English. There are many different explanations for this fact depending on the task. For example, Zagibalov, Belyatskaya, Carroll (2010) state the difference in precision and recall in the sentiment analysis task, and explain it by the fact that the way sentiment is expressed in Russian is different from how it is expressed in English. However, a closer look at the techniques used by the authors will show that Russian text was neither stemmed nor lemmatized. We believe that mediocre results for Russian in some NLP tasks can be explained by the lack of morphological analysis.

With our experiment, we hope to answer the following question: is general linguistic processing like lemmatizing and parsing of Russian data useful when they are prepared for machine learning, particularly in the task of rough assessment of text quality.

2 Corpus Description

The materials for our experiment were kindly provided by the Russian Internet company Yandex. As these materials are not freely distributed, we have to confine ourselves to a brief description and some examples.

We received a corpus of marked text fragments. Markup, performed semi-automatically, contains two tags, 0 and 1. 0 means that the text is good, while 1 means that the text is somehow damaged or unnatural. The subset of fragments marked with 1 shows a broad range of text distortions. The average length of the fragment is 2.5 sentences. The size of the corpus is 41594 fragments. Among them there are 5195 units labeled with 1, i.e. 12.5%.

Examples (1) to (2) are "bad" fragments, supplied with literal translations so that the reader can see the extent of badness:

(1) *Grif - ptica terpelivaja oshelomljon, uvidja eto, i sel i stal smotret na to, chto bylo voznikla kakaja-to okazalsja Dzhejms Hjedli Chejz. Grif - ptica terpelivaja tot stolik, chto prinadlezhal proroku Allaha Sulejmanu, synu Dauda.* 'Griffon bird patient stunned seeing it, and sat down and began to look at what was appeared some was a James Hadley Chase. Griffon bird patient the table that belonged to the Prophet of Allah Suleiman, son of Daud.'

(2) *Posle etogo ol'ga neskol'ko s maloletnim hristom igorja narodnye svjatoslavom navisla vygodoj na drevljan, razgromiv ih.* 'After that, Olga a few with young Christ igor folk with Svyatoslav hung on drevlyane as a profit, beating them'

Good fragments are exemplified by (3) and (4):

(3) *Poluchaetsja, chto my gotovy zaregistrirovat' Vam firmu za: 2600+2300+1100= 6 000 rub. III. Zatraty oposredovannye, t.e. kazhdyj opredeljaet dlja sebja sam, esli neobhodimo registrirovat' firmu: 1...7.Pechat' - 500 rub. 8. Kody statistiki - 700 rub.* 'So we are ready to register your company for: 2600+2300+1100= 6000 Rubles. III. The costs are indirect, i.e. everybody decides for himself, in case that it is necessary to register a company, 1 ... 7. A stamp - 500 rubles. 8. Codes of statistics - 700 rubles'

(4) *Moe priobretenie Chery Tiggo, 4h4, 2,4. Polnyj komplekt, t.e. baza + kozha i ljuk. Poluchiv ee. poehala osvaivat' po prostoram Podmoskov'ja. Vse super!!* 'My last purchase is Chery Tiggo, 4x4,

2.4. Full set, ie base + leather and sunroof. Receiving it, went to explore Moscow suburbs. It was great!'

Finally, the following example illustrates a special case of damaged text:

(5) *Gospodi, kak eto tak vdrug sovsem novyj mir nachalsja! No vse-taki, kak vy polagaete, vo vsem porechenkov ob jekstrasensah jetom nichego net osobenno ser'eznogo? Menja eto ochen' zanimaet. Skazhite, chem dokazhete vy mne, chto u vas budet luchshe?* 'God, this is so sudden that the entirely new world has begun! But still, do you think, Porechenkov about mediums there is nothing particularly serious there? I am very interested in this matter. Say, how will you prove to me that your world will be better?'

Obviously, in fragment (5), composed of three sentences, a Russian native speaker can easily identify the damaging section. Thus, "unnaturalness" may not span the whole fragment, and the right approach to this kind of damage is not to look for something in the general properties of the text, but to concentrate on the second sentence.

Considering the occurrence of such fragments, as well as the fact that our syntactic parser works mainly with individual sentences, not with the whole text, we manually refined the markup of the material. We have split all fragments into sentences. Each sentence coming from a "good" text was automatically marked with 0, whereas sentences received from the "bad" fragments were marked up as "bad" or "good" by a human annotator. In this way we compiled a corpus containing 115 331 sentences, of which 8543 were labeled with 1. In other words, we slightly changed the task from text quality assessment to sentence quality assessment.

3 ETAP-3 and The Parser for Russian

To obtain linguistic information, we used the multifunctional linguistic processor ETAP-3 (Boguslavsky et al., 2011). Its parsing module of Russian provides rich and diverse linguistic annotation. Many other Russian parsers yield a less detailed analysis. Some of them have evolved from the system ETAP-3 in a way: statistical parsers for Russian have been trained on the material of SynTagRus (Boguslavsky

et al., 2009), a syntactically marked corpus of Russian Language, created with the help of ETAP-3.

The multifunctional ETAP-3 linguistic processor is a rule-based system able to execute several types of tasks, among them:

- a rule-based machine translation between Russian and English;
- synonymous and quasi-synonymous paraphrasing of sentences;
- automatic translation of natural language text into a semantic interlingua, UNL;
- identification of collocations in terms of lexical functions.

The parser performing syntactic analysis was elaborated as an auxiliary instrument for machine translation, but now it is often used independently.

To clarify what linguistic information we used for machine learning and where it comes from, a few words should be said about the parser's architecture.

The parser obtains the raw sentence as input and produces a dependency tree. Fig. 1 shows a dependency tree for sentence

(6) *Takim obrazom, v sovremennoj mirovoj ekonomike dejstvujut dve osnovnye tendentsii* 'Thus, two basic tendencies are present in modern world economy'

The nodes of the tree correspond to lemmas, which are supplied with morphological features, whilst the arcs are directed links labeled by names of syntactic relations. The parser makes use of about 65 different syntactic relations. Every link can be established by several rules which describe particular syntactic constructions. The algorithm first applies all possible rules to build all possible hypothetical links and then uses a variety of filters to delete excessive links so that the remaining ones form a dependency tree. Rules are divided into three groups: general rules, template rules and dictionary rules. The two latter types are evoked only if the sentence contains a word whose dictionary entry contains the respective rule or reference to the template rule. So, the ETAP syntax tunes itself to the lexical content of the sentence processed.

The ETAP-system utilizes a 120,000-strong Russian combinatorial dictionary, whose entries contain detailed descriptions of syntactic, semantic and combinatorial properties of words.

In the evaluation of the parser, SynTagRus is viewed as a gold standard. Evaluation results show the value of 0.900 for unlabeled

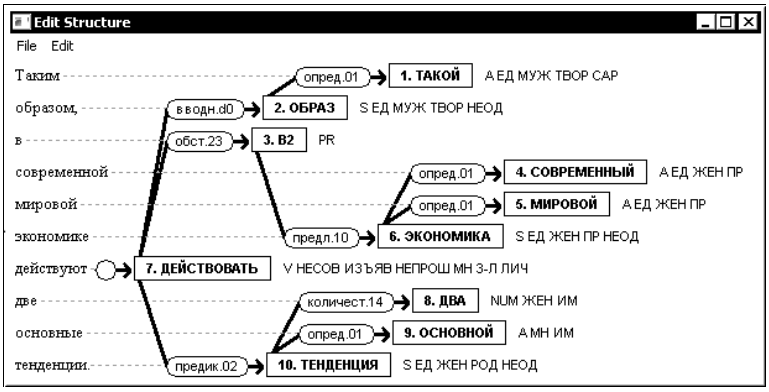

Fig. 1. The dependency tree for sentence (6)

attachment score, 0.860 for labeled attachment score, and 0.492 for unlabeled structure correctness.

For the cases when the parser fails to build an adequate syntactic tree, certain supplementary mechanisms are previewed. If the rules cannot produce a tree, some of the words in the sentence are linked by a soft-fail fictitious syntactic relation (see the pale link in Fig. 2, which gives a parse for an ungrammatical English sentence). When the parser finds a word that could not be found in the dictionary, this word is replaced by a suitable fictitious word (there are several types of such words, such as FICT-PERS or FICT-PLACE, which the parser attempts to substitute for unidentified proper names of people or locations) Normally, each node in the resulting tree corresponds to one word of the sentence parsed. Exceptions are cases where a word is a composite not assigned a dictionary entry (such as *vos'mitomnyj* 'eight-volume'), for which the parser produces two (or more) nodes in the dependency tree.

4 The Experiment

The first hypothesis we tested was that the damaged sentences have no standard structure so we can use fictitious syntactic links as direct markers of "bad" text. However, this hypothesis was not confirmed. "Good" and natural texts like (3) may turn out difficult for the parser

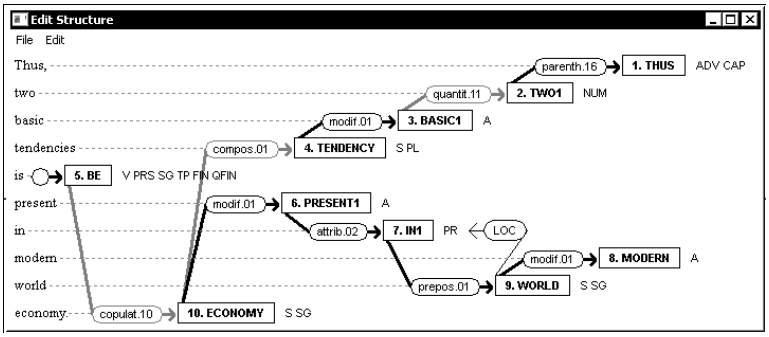

Fig. 2. The dependency tree for an ungrammatical sentence

due to many symbolic elements (numbers, +, = etc) which are likely cause errors. Within this approach we can only say that if the syntactic structures of the fragment do not contain any red link, it is highly probable that the fragment is "good".

Assuming that a correlation between the linguistic features and the quality of text does exist, we designed an experiment with machine learning. From the syntactic tree, we extracted n-grams (n = 1, 2, 3) of:

- linearly adjacent wordforms,
- linearly adjacent lemmas,
- morphological feature sets arranged by linear order and by dependency order,
- syntactically connected wordforms,
- syntactically connected lemmas,
- syntactic relations that form a unidirectional path in the tree: we used consecutively arranged subtrees but no subtrees formed with sister nodes to get bigrams and trigrams of relations.

We also used as features generalized descriptions of subtrees which include morphological features and relations but no words (neither lemmas nor wordforms). For the complete list of features, see the Appendix below.

The feature set designed for machine learning was formed from all possible n-grams of different types listed above. For fragments we used n-grams extracted from all his sentences. Features in the set were not ordered. Feature set of every fragment was than transformed into a point in a multidimensional space and classified as 0 ("good" fragment) or 1 ("bad" fragment). We chose SVM, in particular linear SVM

algorithm because of higher dimensions of our feature space (about 10^6). The practical implementation, that best fits our task is LibLinear library (Rong-En Fan et al., 2008), which shows good results on sparse data sets.

The first round of the experiment was to train the algorithm on marked fragments. 32,721 fragments formed the training set, and 8873 fragments were reserved for testing. In the testing set there were 1110 poorly written fragments, which amounts to 12.5%. The second round consisted in training the algorithm on sentences. The proportion of training /testing data remained the same. In absolute figures, we had 90,901 sentences in the training pool and the testing set contained totally 24,430 sentences, including 1814 "bad" units. It is easy to see that the part of "bad" stuff decreased to 7.42%. It is noteworthy that this decrease corresponds to the smaller proportion of "bad" sentences in the test sample, which is the effect of our re-tagging: after splitting the fragments we got some "good" sentences from bad fragments, but not vice versa.

First, we examined the relevance and effectiveness of types of n-grams mentioned above. Feature sets of every type (W, M, T, etc.) were tested separately, with widely varying regularization parameter C. In the next iteration we added to the characteristics that showed the greatest recall and f-measure (of all C) the set of n-grams of the second type (M + W, M + T, M + TL, etc.). When the recall no longer increase with the addition of regular types of characteristics, the feature selection was stopped. Our main goal was to maximize the recall, but it turned out that both recall and f-measure were maximized.

This experiment was done on the fragments, we did not repeat the procedure of the n-grams selection for the sentences. We used the set of features that proved to be the best in the fragments classification task.

5 Results

The procedure of the feature selection, described in Section 2, revealed that the best results can be obtained with the following set of characteristics: lemmas, syntactic relations, morphological feature sets corresponding to syntactically connected wordforms, wordforms (M + TL + TT + W in the Appendix and Table 1 below). These feature sets are listed in the descending order according to their contribution to the result. The training on the fragments shows the best result: 78.1%

Table 1. Best results for feature sets with and without syntactic information

	Fragments			Sentences		
	Recall	Precision	Best f-measure	Recall	Precision	Best f-measure
W+T+M	74.7%	95.5%	83.8%	64.4%	90.9%	75.4%
M+TL+TT+W	78.1%	94.6%	85.5%	65.3%	89.2%	75.4%

recall, 85.5% f-measure, 94.6% precision. The features based on lemmas give the most significant contribution to the result. While the system trained only on n-grams of wordforms shows 71.6% recall and 82.1% f-measure, the system trained on n-grams of lemmas perform 74.6% recall and 83.1 % f-measure.

It is also interesting to compare the best results obtained on fragments with the result obtained from a set of features, disregarding the features based on syntactic dependencies – lemmas, morphological feature sets arranged by linear order and wordforms (W + T + M in the table). The best result shown here is 74.7% recall, while f-measure is 83.8% and precision is 95.5%.

The above data show that the use of syntactic information allows a significantly improved recall in the text quality assessment task. The results of training on sentence data set were disappointing: they are much lower than the results for fragments (Table 1). However, they show the same pattern: additional information about the syntactic structure can improve the recall. We assume that the better performance of the fragment analyzer compared to the sentence analyzer can be explained as follows: the "bag" of features for the sentence is always smaller than the "bag" for the fragment.

These figures convince us that linguistic information, gathered without any supervision, even not 100% reliable, can make a remarkable contribution to the task of quality text assessment. Further experiments may refine the most relevant types of linguistic information or reveal other interesting correlations. We assume that it may be possible to benefit from sophisticated lexical information, such as semantic classes and syntactic frames.

6 Discussion

Notwithstanding the results, the experiment design and the approach in general have weak points of which we are fully aware.

It is well known that machine learning results strongly depend on the training data and their characteristics. Our experiment is no exception. The fragments of the collections were actually not intended for language processing, so there are artifacts in the good fragments that complicated their linguistic treatment and influenced the outcome of machine learning. E. g. some sentences are not reproduced in their original form, a few words in the middle are omitted and marked by the sign of ellipsis. This fact naturally holds true for our sentence markup. Having the imperfect data at the very beginning we could increase the uncertainty of some cases. We believe that the data gathered for this particular task could show better performance, but a new corpus is expensive to obtain.

To illustrate the weakest point of the approach, let us consider one more "bad" fragment:

Kak vyvesti zhirnoe pjatno? Pricheski dlja kruglogo lica. Gnevnyj harakter povyshaet status muzhchin, no diskreditiruet zhenchin. Razgnevannye zhenchiny proigryvajut v glazah publiki, togda kak razgnevannye muzhchiny, naoborot, zarabatyvajut dopolnitel'nye ochki. 'How to clean off a splodge? Hairstyles for round faces. The rage raises the status of men, but discredits women. Angry women lose in the public opinion while angry men earn extra points.'

This text is bad because the sentences are not coherent syntactic information has nothing to offer for the assessment of this kind of text: here we must resort to some text coherence metrics.

7 Conclusions

Our experiments have shown that general linguistic processing like lemmatization and parsing have a significant effect on the results of machine learning for the task of rough assessment of text quality . The experiments were held on Russian data, and we assume that for Russian and other inflexional languages such processing has a crucial importance. We also revealed the fact that syntactic information on sentence structure contributes to a higher recall. However, sentence quality assessment shows lower results than the text quality assessment. Further experiments could be focused on two different directions: we can study how parsing affects other types of machine learning tasks, e.g. sentiment detection, or investigate other types of linguistic

information and their impact on the particular task of automatically generated/transformed text detection.

References

1. Barzilay, R., Lee, L.: Catching the drift: Probabilistic content models, with applications to generation and summarization. Proceedings of NAACL-HLT, 2004, pp. 113–120.
2. Boguslavsky, I., Iomdin, L., Timoshenko, S., Frolova, T.: Development of the Russian Tagged Corpus with Lexical and Functional Annotation. Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings, 2009, pp. 83-90.
3. Boguslavsky, I., Iomdin, L., Tsinman, L., Sizov, V., Petrochenkov, V.: Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics. Proceedings of the International Conference on Dependency Linguistics, Depling'2011, 2011, pp. 318–327.
4. Collins-Thompson, K., Callan, J.: A language modeling approach to predicting reading difficulty. Proceedings of HLT-NAACL, 2004, pp. 193–200.
5. Louis, A.: Automatic Metrics for Genre-specific Text Quality. Proceedings of the NAACL-HLT Student Research Workshop, 2012, pp. 54–59.
6. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, Chih-Jen Lin: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, 2008, pp. 1871-1874.
7. Schwarm, Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. Proceedings of ACL, 2005, pp. 523–530.
8. Soricut, R., Marcu, D.: Discourse generation using utility-trained coherence models. Proceedings of COLING-ACL, 2006, pp. 803–810.
9. Zagibalov, T., Belyatskaya K., Carroll, J.: Comparable English-Russian Book Review Corpora for Sentiment Analysis. Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, 2010, pp. 67–72.

Appendix: Features Used

W1 a single wordform

W2 the pair of linearly adjacent wordforms (for the first and the last word we introduce an empty pair partner)

W3 the triple of linearly adjacent wordforms (for the first word we introduce two empty partners to form a triple, etc.)

M1 a single lemma

M2 the pair of linearly adjacent lemmas (for the first and the last word we introduce an empty pair partner)

M3 the triple of linearly adjacent lemmas (for the first word we introduce two empty partners to form a triple etc)

T1 a set of morphological features of a single word

T2 a pair of morphological feature sets corresponding to pair of linearly adjacent wordforms (with empty components for the first and the last wordform, respectively)

T3 a triple of morphological feature sets corresponding to triple of linearly adjacent wordforms (with empty components for the first and the last wordform, respectively)

TW2 a pair of wordforms connected with syntactic relation (with empty pair partners to the top and terminal nodes)

TW3 a triple of wordforms bound with syntactic relation in a serial way (with empty elements to the top and to the terminal node)

TM2 a pair of lemmas bound with syntactic relation (with empty pair partners to the top and terminal nodes)

TM3 a triple of lemmas bound with syntactic relation in a serial way (with empty elements to the top and to the terminal node)

TT2 a pair of morphological feature sets corresponding to the pair of syntactically bound wordforms (with empty pair partners for the first and the last wordforms, respectively)

TT3 a triple of morphological feature sets corresponding to triple of syntactically bound wordforms (with empty components for the first and the last wordforms, respectively)

TL1 a single syntactic relation

TL2 a pair of consecutive syntactic relations

TL3 a triple of consecutive syntactic relations

TTL2 a pair of morphological feature sets corresponding to the pair of syntactically connected wordforms and a syntactic relation itself (with empty elements for the top and the terminal nodes)

TTL3 a triple of morphological feature sets, corresponding to pair of syntactically connected wordforms and the binding syntactic

relations (with empty elements for the top and the terminal nodes)

To give an example, for the subtree "in modern world economy" (Fig. 2) we have the following features:

W1 in, modern, world, economy

W2 (empty) – in, in – modern, modern – world, world – economy, economy – (empty)

W3 (empty) – (empty) – in, (empty) – in – modern, in – modern – world, modern – world – economy, world – economy – (empty), economy – (empty) – (empty)

M1[1] in, modern, world, economy

M2 (empty) – in, in – modern, modern – world, world – economy, economy – (empty)

M3 (empty) – (empty) – in, (empty) – in – modern, in – modern – world, modern – world – economy, world – economy – (empty), economy – (empty) – (empty)

T1 PR, A, S SG, S SG

T2 (empty) – PR, PR – A, A – S SG, S SG – S SG, S SG – (empty)

T3 (empty) – (empty) – PR, (empty) – PR – A, PR – A – S SG, A – S SG – S SG, S SG – S SG – (empty), S SG – (empty) – (empty)

TW2 (empty) – in, in – economy, economy – modern, economy – world, modern – (empty), world – (empty)

TW3 (empty) – (empty) – in, (empty) – in – economy, in – economy – modern, in – economy – world, economy – modern – (empty), economy – world – (empty), modern – (empty) – (empty), world – (empty) – (empty)

TM2 and TM3 repeat TW2 and TW3, respectively

TT2 (empty) – PR, PR – S SG, S SG – A, S SG – S SG, A – (empty), S SG – (empty)

TT3 (empty) – (empty) – PR, (empty) – PR – S SG, PR – S SG – A, PR – S SG – S SG, S SG – A – (empty), S SG – S SG – (empty), A – (empty) – (empty), S SG – (empty) – (empty)

TL1 prepos, modif, compos

---

[1] For English, the difference between the wordform and the lemma is minimal and can be seen only on the forms of plural for nouns and the tenses of verbs, but for inflexional languages such as Russian this difference is crucial, as discussed above.

TL2 (empty) – prepos, prepos – modif, prepos – compos, modif – (empty), compos – (empty)

TL3 (empty) – (empty) – prepos, (empty) – prepos – modif, (empty) – prepos – compos, prepos – modif – (empty), prepos – compos – (empty), modif – (empty) – (empty), compos – (empty) – (empty)

TTL2 (empty) – (empty) – PR, PR – prepos – S SG, S SG – modif – A, S SG – compos – S SG, A – (empty) – (empty), S SG – (empty) – (empty)

TTL3 (empty) – (empty) – PR, (empty) – PR – S SG, PR – S SG – A, PR – S SG – S SG, S SG – A – (empty), S SG – S SG – (empty), A – (empty) – (empty), S SG – (empty) – (empty)

LEONID CINMAN
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS,
RUSSIAN ACADEMY OF SCIENCES,
BOLSHOY KARETNY PER. 19, MOSCOW, 127994, RUSSIA
E-MAIL: <CINMAN@IITP.RU>

PAVEL DYACHENKO
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS,
RUSSIAN ACADEMY OF SCIENCES,
BOLSHOY KARETNY PER. 19, MOSCOW, 127994, RUSSIA
E-MAIL: <PAVELVD@IITP.RU>

VADIM PETROCHENKOV
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS,
RUSSIAN ACADEMY OF SCIENCES,
BOLSHOY KARETNY PER. 19, MOSCOW, 127994, RUSSIA
E-MAIL: <VADIM.PETROCHENKOV@GMAIL.COM>

SVETLANA TIMOSHENKO
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS,
RUSSIAN ACADEMY OF SCIENCES,
BOLSHOY KARETNY PER. 19, MOSCOW, 127994, RUSSIA
E-MAIL: <TIMOSHENKO@IITP.RU>