ПОСТРОЕНИЕ СИСТЕМЫ ГЕНЕРАЦИИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ: ПЕРВИЧНАЯ РАЗРАБОТКА

Клименченко Л. В. (lyu.klimenchenko@gmail.com),
Пискунов В. И. (vl.i.piskunov@gmail.com)

НИУ Высшая школа экономики

Ключевые слова: компьютерная лингвистика, генерация текстов на естественном языке, генерация новостей, русский язык

BUILDING NATURAL LANGUAGE GENERATION SYSTEM IN RUSSIAN: PRIMARY REALIZATION

Klimenchenko L. V. (lyu.klimenchenko@gmail.com),
Piskunov V. I. (vl.i.piskunov@gmail.com)

National Research University Higher School of Economics, Moscow, Russia

The research in natural language generation (NLG) in Russia is not very developed and advanced nowadays. There are only few known Russian projects in this field.

In this paper we present a system that is being created for generating stock news in Russian. The project is set up by request of RBC media-group and is still under development.

Here we give a short overview of a similar text generator realized for English and describe the first version of our NLG system. Currently the system is pattern-based and uses time-referenced database as input data. In the paper we describe the architecture of the system, the rules of filling in the sentence patterns and combining them within the text structure, give an analysis of the lexical material data and present some output texts. However, we realize all disadvantages of the pattern-based approach and as a result we make suggestions for possible further reformation and implementation of our system.

Key words: computational linguistics, natural language generation, news generation, Russian

1. Natural Language Generation: State of Art

Natural language generation originated as subfield of computer linguistics several decades ago. In 80–90s it was actively developing in the USA, Canada and European countries, see [Reiter 1995], [Reiter, Dale 2000], [McKeown 1982]. The task of NLG systems is to generate a text in natural language from some non-linguistic source such as databases, semantic representations etc. To accomplish this one a considerable amount of knowledge in various fields is needed: domain knowledge, linguistic knowledge including pragmatics, semantics, syntax, morphology, phonology, etc. [Smedt, Horacek, Zock 1996].

There are some major approaches to NLG. Pattern approach is quite simple, robust and rather popular. Pattern-based NLG systems work with language information encoded as strings of symbols. Its main feature is in combining ready-made expressions and word forms with strings with gaps, which helps to avoid ungrammatical structures in output text, but also narrows their variety. Another possible approach is language motivated systems, that work with non-linguistic data like databases, knowledge bases, formal languages and semantic representations. Language motivated systems base on using text characteristics and linguistic knowledge (formal grammars, anaphora etc.). For solving problems developers of language motivated systems sometimes use pattern-based approach [Sokolova, Boldasov 2004].

Here is a brief review of some well-known NLG Systems.

ANA (1983)—generation of stock reports in English (for more information see the next section).

FoG (1994)—generates forecast reports for ships based on measures of atmosphere parameters (such as wind direction, wind strength, temperature etc.).

AGILE (2000)—generation of software manuals in several languages (Bulgarian, Czech and Russian).

Komet Project MultiLinual (1990–2000s)—special source for text generation in different languages It can work with English, French, German, Dutch, Japanese, Czech, Bulgarian and Russian data. The system is based on the grammar written with the framework of Systemic-Functional Linguistics (SFL).

DEMlinG (2004)—another multilingual system, that operates with interlinguistic representations for further text generation. It gets data from database as an input, modifies it into morphosyntactic structures and then produces the text. DEMlinG has an integrated Russian morphology component that makes texts generation in Russian possible for the system. [Boldasov 2003]

ARRIA (present)—the one of the most famous commercial NLG system that was developed by Ehud Reiter and Robert Dale, the authors of [Reiter, Dale 2000], which is the milestone work in the NLG area. The aim of ARRIA is to perform a sophisticated analysis of big data and huge datasets with a logical conclusion and generate an articulate and grammatically correct natural language summary. As it is a commercial project, the company does not reveal the methods that are implemented in this system.

There is also one research that should be mentioned here, namely the system for automatic generation of sports commentaries that was implemented on Russian material [Tokareva, Bolshakova, Bordachenkova 2006]. The system is pattern-based

and also involves operation with so called event types. Its difference from our system is that the commentaries generated by it are sentences while our system gives short texts as an output.

2. A system for generating natural language stock news

The stock quotes change considerably during the day, which means that financial media should react quickly to each remarkable change and announce the news very often and very fast. That is why it becomes necessary to make a system that receives the information about latest changes and generates short news on the base of it.

In 80s this idea was implemented by Karen Kukich in Ana system [Kukich 1983]. Ana is a knowledge-based report generator, its main tasks are inferring semantic messages from the data in database, mapping those messages into phrases using the phrasal lexicon and ordering them according to the rules of clause-combining grammar and rhetoric constraints [Kukich 1983: 146]. Ana system contains four modules: fact generator, message generator, discourse generator and text generator.

The text generator does the most complicated processing as regards lexicon selection, morphology, anaphora, syntax, punctuation, and control of discourse mechanics. In example (1) there is a text generated by Ana.

(1) *After climbing steadily most of the morning, the stock market was pushed downhill late in the day. Stock prices posted a small loss, with the indexes turning in a mixed showing yesterday in brisk trading. The Dow Jones average of 30 industrials surrendered a 16.28 gain at 4pm and declined slightly, finishing the day at 1083.61, off 0.18 points.* [Kukich 1983: 146]

Our project also has a task of creating NLG system that generates natural language stock news daily. The difference between Ana and our project is that first, our system is developed for generating texts in Russian and second, the output texts that it generates are of four types according to the time when they are generated and the kind of stock change. For primary realization of the system we use time-referenced database of stock quotes as an input data source and rule-based pattern approach.

3. Preliminary work

3.1. Developing text models

As it was mentioned earlier, our system is supposed to generate texts of four different types. The type of the text depends on the characteristics of the event at the stock market and the kind of the subject the text is generated about, either stock indexes or securities.

The text types are the following:

Morning—The description of the last trading day (generally and particular) and morning and how indices have performed during the first minutes after trades started.

Day Review: How indices have performed during the current trading day.

description in one sentence.

Big News (for MOEX and/or RTSI indexes): The description of the significant change that has been experienced by the index recently.

Big Stock News: The description of the significant change that has been experienced by the security recently.

We define two types of events: first type is a situation when there may not be any considerable change at the stock market, and we make a brief review of it; and the second is an observation of considerable index quote/share price change at the stock and we consider it as a basis for breaking news.

Our system distinguishes two types of the subjects in text: stock indexes (such as MOEX and RTSI) and company shares. That can be explained by two facts. Firstly these two entities have different nature: while stock index is a measurement of the value of the stock market that is computed by the prices of the selected stock shares, by saying shares we refer to securities of some company or corporation that are traded on the stock market. Secondly they appeared to have quite different behavior that is reflected in numerous different lexical expressions of these entities (See section 3.3 for more information).

While analyzing the RBC archive of news, it became clear that the output texts should contain mostly stock index reports as they appear everyday two or three times, while security news can be encountered only if their rate has changed dramatically. This information demand of RBC should have been also taken into account and it approved the following classification of text types. After the analysis of the archive a corpus of target text was created. Its volume is relatively small and consists of 100 short news and abstracts from some analytic articles that also contain our target texts.

The other very important feature that appeared to be clear-cut is that all four models use the same structure, either within one model or through all the text types, and the lexical variety in them is narrow. This might justify our intention to build a template-based NLG system.

As we concentrate in the template-based approach, the information structure of the texts does not have deep discourse analysis.

The information structure of the morning news text model.

Morning News

1) Trades beginning: How indices have performed during the first minutes after trades started (general tendency)
2) The performance of particular indices
3) The characterization of the last trading day (generally and particular)
4) The determining of the trade value for the previous trading day.
5) The list of securities that have shown the most significant changes during the trading day

The first step is to distinguish messages in every model. For now the complex text structure is avoided, so any message is converted only into one sentence.

3.2. Learning how to speak "Financian" or collecting the lexicon

The next step after creating discourse plans for the text models is collecting the lexicon. It is one of the most important parts of work, because according to Becker, who introduced the notion of the lexicon, people generate utterances 'mostly by stitching together swatches of text that they have heard before'. [Becker 1975: 70–73]

For this purpose the target text corpus was used. However, we decided to extend our corpus to work with bigger range of vocabulary. The collection was enlarged by adding some stock market articles and reports from other internet based economic media sources (finam.ru, ivestcafe.ru and lenta.ru). The current corpus consists of about 200 texts.

To provide the output texts with language variation, we carried out the analysis of lexis. The variation itself is very important in Russian stylistics and rhetorics, so the generated text with no repetitions should look more like human written text. As a result of lexical analysis the three main groups of the vocabulary were determined: subjects, predicates with attributes and time expressions.

Since we distinguish two types of subjects: indices and securities, for each subject the list of possible synonyms was collected to provide our system with more options of lexical variation. However, the extension of the vocabulary was not the only goal. The co-occurrence with the verb attributes was also analyzed during this section, to be used in the further work with predicates. The most important part of lexical analysis is the evaluation of predicates that describe the behavior of indices and securities. The vast range of the vocabulary was collected, analyzed and sorted in this section. Time expressions can be classified in four groups: opening ('in the morning', 'at the beginning of the trading day', etc.), duration ('during the trading day', etc.), day reference ('today', 'yesterday', etc.) and exact time expressions. The using of these three lexical groups depends on the type of the chosen text model.

More information about lexical analysis will be given in section 4 of this article.

4. System Architecture and implementation

The system is implemented in Python and consists of three independent, sequential components: database to text model and event types (facts) conductor (calculating stage), lexicalizer (conceptual stage), text constructor (text generation stage). At the first stage the system gets data from the database as an input and calculates the text model type and the event types. The database provides the information about quotes and their changes such as quotes at the opening and at close of exchange, last change in points, last change in percent, the lowest and highest index etc.

Table 1. Input data from database

OPEN	CURR	CHPT	CHPRC	CHOPENP	CHOPENPC	HIGH	LOW
1654.9	1696.0	40.0	2.4	41.0	2.5	1698.8	1642.5
1654.9	1696.0	40.0	2.4	41.0	2.5	1698.8	1642.5

Legend: *OPEN*—a quote of an index at the opening of the trading day; *CURR*—a current quote of an index; *CHPT*—a change to the last significant mark in points; *CHPRC*—a change to the last significant mark in percent; *CHOPENP*—a change to the open value in points; *CHOPENPC*—a change to the open value in percent; *HIGH*—the highest mark during the trading day; *LOW*—the lowest mark during the trading day.

After analyzing this data the system chooses the text model and gives the event types as output. There are two basic types of events 'plus' and 'minus', but we also distinguish so called 'super-plus' and 'super-minus', when the changes are considerable. The quotes can display unstable behavior during the trades, for example to grow up in the morning and sink by the evening, therefore one can call this 'compound events'. If we take the binary ones (plus—minus, super-plus—minus), there are 16 possible combinations. While working with target corpus we found examples (sentences illustrating such events) only for seven combinations.

Table 2. Event types presence in target corpora texts

Plus-Plus	Plus-Minus	Plus-SuperPlus	Plus-SuperMinus
Minus-Plus	Minus-Minus	Minus-SuperPlus	Minus-SuperMinus
SuperPlus-Plus	SuperPlus-Minus	SuperPlus-SuperPlus	SuperPlus-SuperMinus
SuperMinus-Plus	SuperMinus-Minus	SuperMinus-SuperPlus	SuperMinus-SuperMinus

Leaping ahead we should note that the constituents of the compound event can be placed in reversed order in their lexical expression, i. e. SuperMinus-Plus can be expressed as 'the quotes began to rise after a significant fall'. According to our corpora observations it must be a tendency to place the 'super' event after the normal one.

The event types serve as an input for lexicalizer, a part of the system which main task is to provide it with adequate lexical variants for the chosen events and a text model.

After the analysis of semantic features of subjects and their lexical realization the matrix of possible attributes was created. The following table displays generally acceptable lexical attributes for the words 'indexes' and 'securities' in the Russian language:

Table 3. The Acceptable attributes of the subjects

	change Percent	**change Quote**	**change Price**	**hasEdge Quote**	**hasEdge Price**
Index	+	+		+	
Security	+		+		+

Legend: *changePercent*—a value of the subject change can be measured in percents; *changeQuote*—a value of the subject change can be measured in points; *changePrice*—a value of the subject change can be measured in some currency; *hasEdgeQuote*—a subject can be measured in points; *hasEdgePrice*—a subject can be measured in some currency.

Then we analyzed predicates and allocated them into three semantically motivated groups: predicates that describe positive tendency of index/security change; the ones that describe negative tendency of the assets; and the vocabulary used only to state the value of index/security. The last group is relatively small as its lexical variety is quite narrow. Afterwards predicates were analyzed according to the attributes that may co-occur with them. For each predicate we built the matrix of all possible attributes:

Table 4. The possible attributes for the predicates

	change PerCent	**change Quote**	**change Price**	**hasEdge Quote**	**hasEdge Price**
Predicate_1	+	+	+	+	+
Predicate_2	+		+		+

Legend: *changePercent*—a predicate can describe a change of some entity in percent; *changeQuote*—a predicate can describe a change of some entity in points; *changePrice*—a predicate can describe a change of some entity in some currency; *hasEdgeQuote*—a predicate can settle an edge in points reached by some entity; *hasEdgePrice*—a predicate can settle an edge in some currency reached by some entity.

The intersection of this matrix with one of the subjects helps to choose the appropriate predicate for the given subject. Some predicates are used only in some text models. They can be called 'final tendency predicates' as they generally characterize the behavior of the assets during the trading day *(закрыться в красной зоне* 'finish in red zone'). For that reason they are used only in the *Morning* and *Day Review* text models. Also these predicates might demand some special subjects that can be used only with them within these models (*торги* 'trades' and others).

When the lexis is chosen the only one step remains left, namely to put the words in right order. As far as we do not have any grammar module and the sentences of the output texts have fixed structure (because of pattern-approach), the words are being inserted in exact grammatical forms, which suit the sentence structure and do not violate grammar rules. There are some examples of the output texts.

(2) Generated text in Russian

Сегодня торги проходили в красной зоне. Утром индекс ММВБ начал торги понижением и продолжал устремляться вниз. В то же время рухнув утром, индекс РТС продолжил сильное понижение. Так индекс ММВБ понизился на 0.39% до отметки в 1748 пунктов, а индекс РТС — снизился на 4.4% и достиг 804 пунктов. Объем торгов по итогам дня составил 700 миллионов долларов США.

English translation

Today the trades ran in the red zone. In the morning index MOEX started to reduce and to fall continuously. At the same time RTSI index proceeded with its lowering. So MOEX lost 0.39% and made up 1748 points, RTSI fell to 4.4% and reached the grade of 804 points. The volume of the trading section was 700 millions of dollars.

(3) Generated text in Russian

После открытия торговой сессии российские индексы начали расти. Днем индекс ММВБ подрос на 1.05 % до 1774 пунктов, а индекс РТС поднялся на 3.33 % до 854 пунктов. Вчера торги на российском рынке акций завершились повышением. Так индекс ММВБ поднялся на 3.45 % до отметки в 1755 пунктов. Другой основной российский индекс — РТС — увеличился на 2.61 % и достиг 826 пунктов. Объем торгов по итогам дня составил 695 миллионов долларов США.

English translation

After the trades opening Russian indexes began to grow up. During the day MOEX rose for 1.05 % to 1774 points, RTSI climbed 3.33 % to 854 points. Yesterday the trades finished with raising. So MOEX rose for 3.45 % to the grade of 1755 points. The other Russian index RTSI increased by 2.61 % and made up 826 points. The volume of the trading section was 695 millions of dollars.

To provide some variability for the text generation our templates consist mainly of slots where the words from lexical groups should be inserted, but also contain few fixed words. The full template for the latter text looks as following:

{initial_time} {subject} начали {predicate}. {particular_time} индекс ММВБ {predicate} на {%value} пунктов или {%value} % до {%value} пункта, а индекс РТС {predicate} на {%value} % до {%value} пункта. {past_time} торги на российском рынке акций {predicate} {attribute}. Так индекс ММВБ {predicate} на {%value}% до отметки в {%value} пунктов. Другой основной российский индекс — РТС — {predicate} на {%value} % и достиг {%value} пунктов. 'Объем торгов по итогам дня составил {%value} миллионов долларов США.

It is also very important to mention that we have several models for each sentence. For example we have these models for the first sentence in the text:

{initial_time} {subject} начали {predicate}.
Торги на российском рынке акций {predicate} {attribute} основных индексов.
Основные индексы {predicate} день {attribute}.

Having one of the possible variants chosen, the rest of the text is generated taking it into account to perform the best coherence possible. We believe that this approach provides the maximum variability for the templates.

At the text generation stage the system fills in the gaps in patterns with the words from their groups. Each lexical group is either semantically motivated, as it is described above, or semantically and grammatically motivated. For example words падать and падение "fall" both belong to the minus group semantically, but as they are different parts of speech we have to distinguish two groups: 'minus_verb' and 'minus_noun'. For each gap in the pattern there is a restricted number of suitable lexical groups. When the text is being generated the program randomly chooses the words

of these lexical groups and inserts them into the pattern. Since the first choice was made lexical restrictions start to apply. These restrictions are realized by some simple algorithms that are based on the information described above. They play an important role in our system, as they are essential for generating coherent texts.

This method works quite well, except the situations where we face the problem of tautology. Unfortunately by now we failed to fix it in all such cases, especially when it is a question of same-root verbs and nouns that are used to express the event type.

5. Conclusion

In the paper we described the system for generating stock market news in Russian in its primary state of development. Pattern-based approach is quite easy and serves well for getting adequate results. Nevertheless it is not that good for generating longer and more complex texts, as it narrows the scope of possible lexical realizations. Another lack that should be fulfilled in the next version of our system is the absence of grammar module. The fact that Russian language has very rich and developed morphology complicates the implementation of the pattern approach.

By now we should consider that we advanced quite a lot in building NLG system in Russian and found out the disadvantages of the pattern approach. Therefore the next task is going to be fixing failures and improving the process of text generation as a whole.

References

1. *Becker J.* (1975), 'The Phrasal Lexicon' // Theoretical Issues in Natural Language Processing, ed. B. I. Nash-Webber, Cambridge, Massachusetts, pp. 70–73
2. *Boldasov M. V.* (2003), Paradigms of Natural Language Text Generation in DEMlinG [Paradigmy generacii EYa tekstov v instrumental'noy srede DEMlinG] // Komp'uternaya lingvistika i intellectual'nye tekhnologii. Trudy Mezhdunarodnoy konferentsyi Dialog'2003, M: Nauka, 2003, s. 66–75
3. *Kukich K.* (1983) Design of a Knowledge-Based Report Generator // Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, pp. 145–150.
4. *McKeown K. R.* (1982), "The TEXT System for Natural Language Generation: An Overview" // Proceedings of the Twentieth Annual Meeting of the Association for Computational Linguistics, Toronto, Canada, 113–120
5. *Smedt K., Horacek H., Zock M.* (1996), Architectures for Natural Language Generation: Problems and Perspectives // Adorni G. & Zock M. (Eds.), Trends in natural language generation: An artificial intelligence perspective (Springer lecture notes in artificial intelligence 1036), Berlin: Springer, pp. 17–46.
6. *Reiter, E., Dale R.* (2000), Building Natural Language Generation Systems, Cambridge University Press

7. *Sokolova E. G., Boldasov M. V.* (2004), Natural Language Text Generation (state of art) [Avtomaticheskaya generatsyya tekstov na EYA (portret napravleniya)] // Komp'uternaya lingvistika i intellectual'nye tekhnologii. Trudy Mezhdunarodnoy konferentsyi Dialog'2004, M: Nauka, 2004, s. 565–572
8. *Tokareva M. Yu., Bolshakova E. I., Bordachenkova E. A.* (2006), Automatic Generation of the Sport Commentary [Avtomaticheskaya generatsyya sportivnogo kommentariya] // Komp'uternaya lingvistika i intellectual'nye tekhnologii. Trudy Mezhdunarodnoy konferentsyi Dialog'2006, M: Nauka, 2006, s. 498–503
9. *ARRIA* — http://www.arria.com
10. *KPML* — http://purl.org/net/kpml