ИЗВЛЕЧЕНИЕ ФАКТОВ ОБ ОТНОШЕНИЯХ МЕЖДУ ПЕРСОНАЖАМИ ИЗ ХУДОЖЕСТВЕННЫХ ТЕКСТОВ

Бодрова А. А. (anastasie.bodrova@gmail.com),
Бочаров В. В. (victor.bocharov@gmail.com)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: извлечение фактов, художественная литература, метод лексико-синтаксических шаблонов, Томита-парсер

RELATIONSHIP EXTRACTION FROM LITERARY FICTION

Bodrova A. A. (anastasie.bodrova@gmail.com),
Bocharov V. V. (victor.bocharov@gmail.com)

Saint-Petersburg State University, Saint-Petersburg, Russia

Our paper describes a method for relationship extraction from Russian-language literary fiction, it is expressed in applying lexico-syntactic patterns. The method was used for extracting facts from formal texts and brought meaningful results. The hypothesis is if the method is suitable for texts, which are non-formal. As an instrument, we chose Tomita-parser, which is based on the formalism of context-free grammars.

For extracting facts about the relations, we used the grammars, written on Tomita language, which describes the way of names extraction, shallow syntax and lexico-syntactic patterns.

We used extracted facts to construct relationship graphs, where characters are vertices and relations between them are edges. As a result, we got about 30 lexico-syntactic patterns, which can extract a set of facts about relations from the literary fiction text.

Keywords: relation extraction, literary fiction, lexico-syntactic patterns, Tomita-parser

1. Introduction

As it is commonly held, literature reflects the life and structure of the social world around. So in its base text keeps the structure of interactions between characters and objects of the plot like in a reality.

The net of the relations is one of the ways of text conceptualization. This picture shows not only connections between characters, but also emphasizes key personages and meaningful clusters of interactions.

Such type of extracted data can be used for literary analysis, for filling databases (e.g. Freebase[1], DBPedia[2]) or for information retrieval (e.g. Amazon.com[3]).

We explored a question of possibility to build a relationship graph for literary fiction, using only lexical and syntactic features and some quantitative characteristics. To realize that idea, we tried a method of lexico-syntactic patterns, which was used predominantly for the formal types of texts before. Our approach includes writing a grammar for Tomita-parser. The grammar consists of parts that are responsible for names extraction, shallow syntax and lexico-syntactic patterns.

Parser extracts information about characters and relations between them. Using retrieved data, we construct a graph, where characters are vertices and relations are edges.

We carried out this work on corpus of Russian-language texts.

In the following sections, we observe related work on conceptualization and relationship extraction for literary fiction. Then we describe our method, ongoing work and the tool used in more detail. After that we present graphs and evaluation received.

2. Related work

Existence of rich literal heritage raises a question about saving it in a convenient form for analysis and studying.

The founder of Stanford Literary Lab, F. Moretti, wrote a book (2007), where he claims that ”literature scholars should stop reading books and start counting, graphing, and mapping them instead”.

This idea was put into effect, when the researchers analyzed big data of literary texts (Jockers, 2013; Michel and Liberman, 2010). With the help of computational analysis, the texts were explored on lexica, linguistic patterns, phrases and meta data. These approaches show the picture of literary history and culture trends.

Apart from looking at big data, text and its plot still remain an important concept of literary analysis.

F. Moretti suggests to use a network of characters to analyze the plot (2010). Since that time, the number of attempts to retrieve information about characters and their relations from literary fiction is increasing.

Set of researches on automated generation of social networks are based on conversational interactions between characters. Firstly a method was presented by Elson et al. (2010). They used quoted speech or dialogues to extract information about who is talking to each other.

---

[1] Available on https://www.freebase.com/

[2] Available on http://dbpedia.org/

[3] Available on http://www.amazon.com/Search-Inside-Book-Books/b?node=10197021

Agarwal (2011) proposed a concept of social events, which are "events that occur between people where at least one person is aware of the other and of the event taking place". He separate two types of interactions: explicit and implicit ones. His proposal enlarges the diapason of possible interactions extracted , and consequently a set of approaches for building a network(Park et al. 2013; Makazhanov et al. 2012; He et al. 2013).

What is more, some researchers add to a network places, where events occur (Marazzato and Sparavigna, 2014; Lee and Yeung, 2012).

Compared to these researches, we firstly concentrated on a network of characters only, added names of relations and focused our method on Russian language.

3. Relationship extraction

In order to realize the aim of building a social network on literary fiction text, we used the method of lexico-syntactic patterns.

3.1. Tool

The choice of an instrument for extracting information from texts for Russian language was among such tools as GATE[4], LSPL[5], AIRE[6], AGFL[7], Tomita-Parser[8]. We selected the last one as more available and user-friendly.

Tomita-parser is an instrument for extracting structured data from the natural language texts. It is based on the algorithm of GLR-parsing[], which uses the formalism of context-free grammars.

3.2. Method

In order to make a graph on literary fiction text, we defined the core concepts in a following way:

Character—the participant of the plot, which has a name.

Relation—1) the connection between characters, which is explicitly expressed ("*Athena, Zeus's daughter*")

2) which is expressed by frequency of joint mentions.

---

[4] Available on http://gate.ac.u

[5] Available on http://lspl.ru/

[6] Information about is available on http://www.linux-ink.ru/static/Docs/NauLinux/School/5.8/slce.58.nauschool.5x/Software-i386/NauLinux/Misc/Extras/Aire/aire.html

[7] Available on http://www.agfl.cs.ru.nl

[8] Documentation is available on http://api.yandex.ru/tomita/doc/tutorial/concept/about.xml

In order to extract information, explicitly shown, we tried to use the method of lexico-syntactical patterns. The point of method is in forming a pattern, which describe syntactic structure, grammar features and key-words vocabularies. While having a pattern, the search inside the text can find all the chains of words, which become relevant results. Firstly proposed by Hearst (1992) for extracting hyponymical relations, this method was effectively used for formal types of texts, e.g. news texts (Yandex.Press-Portraites[9]), biomedical literature (Sugiyama et al.2013).

We tried to apply this method to literary fiction—a non-formal type of texts.

3.3. Data Analysis

We obtained electronic encodings of the texts from free online-library Flibusta[10]. Total it contains more than 250000 books. We made a small representative subcorpus, which included:

- Russian fiction (B. Pasternak "Doctor Givago", A. Chekhov "Dushechka")
- foreign fiction (Y. Yalom "When Nietzsche wept", J. Austen "Pride and Prejudice")
- myths(N. Kun "Legends and Myths of Ancient Greece", J. Tolkien "Silmarillion")
- fairy-tales(A. Volkov "Wizard of Emerald City")
- historical texts("Short story of Romanov dynasty")

These texts are third-person narrative and have a large set of characters with interactions.

The representativeness is reached by the different types of lexica and syntactic structures, inherited to these text types.

After getting texts, they were looked through for defining the features of syntax and lexica. Then we formed key-words vocabularies, expressed roles in relations (мать 'mother', тётя 'aunt', …), relationships names (дружба 'friendship', брак 'marriage', …), actions (любить 'love', воспитывать 'educate', …) and descriptions (женатый 'married', …). For enhancing the vocabularies of relations we used vocabularies of synonims.

3.4. Grammar

For Tomita-parser's work, it is needed to make a grammar, which explicate the information for searching. Our grammar had three parts: names extraction, shallow syntax and lexico-syntactic.

Names Extraction (NE): Names extraction has two steps. Firstly, we made a grammar for detecting non-dictionary names and checking if each of them belongs to a person or a geographical entity. Then we put them into separate vocabularies.

---

[9] Information about available on company.yandex.ru/press_releases/2006/0404/

[10] Available on http://www.flibusta.net

The second step constructs a name from a set of its smaller parts: FirstName, SecondName, Patronymic, Non-Dictionary-Name, and attributes: titles (мистер “mister”, товарищ “camarade”..) and nobiliary particles (де “de”, ля “la”...). Some names were put into the vocabularies with the variants of their diminutive forms. While extracting, all this names are normalized to the unified name. Also, we did a simple coreference resolution, united that names, that one of them is a substring of another. Problem of coreference is still very important, because many cases are lost due to that fact, that resolution needs semantically-oriented methods, that are difficult to put into effect.

Shallow Syntax: The aim of describing such syntactic structures as noun and verb phrase is to improve the search of chains, which suit a pattern. On this level, we should express all the possible subordinators of a noun and a verb and the variants of expression homogenous parts. This description helps to detect a noun (verb) not as a word, but as a head of a noun (verb) phrase.

Lexico-Syntactic Patterns: Lexico-Syntactic Patterns are aimed to extract information for filling the table (Fig. 1):

(1) *Джейн, сестра Элизабет,… “Jane, Elizabeth’s sister, …”*

FPerson	**SPerson**	**Role**
Джейн *Jane*	Элизабет *Elizabeth*	Сестра *Sister*

For writing patterns, we firstly search for the sentences that contains two names and a word from vocabularies, expressing relations, and chose those ones, that bring meaningful information and can be described by a pattern. After that we construct a list of about 30 patterns.

3.5. Quantitative relation extraction

Obviously, characters, that frequently appeared together during the plot, have a connection. For defining it, we formed a list of heroes, in order of their appearing during the story and formed pairs in a distance 2, then the were counted by MI metric. The toppest results were identified as a strong connection between characters. There is no name for such.

3.6. Graphs

For building graphs we used Graphviz[11] the tool for automatic graph visualization, using the description on DOT-language.

All the information, extracted by the parser and quantitative relation extraction, was collected and rewrote on the DOT-language.

[11] Available on http://www.graphviz.org

As a result we derived graphs (Fig. 1), where vertices are submitted by characters and edges by relations between them.

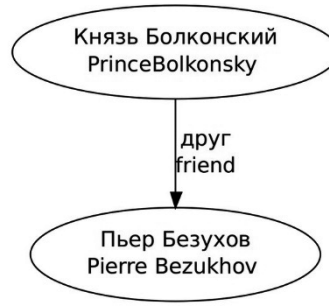


Fig. 1. Example of a graph of relations

4. Evaluation

As a result we derived about 30 lexico-syntactic patterns for extracting characters and relations between them. That method presents more structured and full results for texts with a big amount of characters. But short texts and texts with difficult stylistics were not successful.

Myths and fairy-tales show to be more pattern-based. That happens, because the story of the unreal world needs to make lots of description about its heroes and the genre supports the definite stylistics, but on the other hand, such texts bring problems with coreference due to a big amount of repeated names for different characters. What is more, a lot of meaningful characters are called once or twice, while they are important for the story, but do not take part in a plot, that make their extraction a bit difficult. Graphs, got for myths, are in Appendixes (3, 4).

4.1. Results

We manually annotated text “Pride and Prejudice” for the pairs of characters, whose connection is meaningful for the plot and counted the precision and recall.

By this moment, we compared our results with those, that were made by other researchers, on the novel.

	Precision	**Recall**
Our method	0.67	0.8
He et al.	0.8	?
Makazhanov et al.	1	0.4
Elson et al.	0.9	0.5

The graph for this novel is presented in Apendix 1.

5. Conclusion

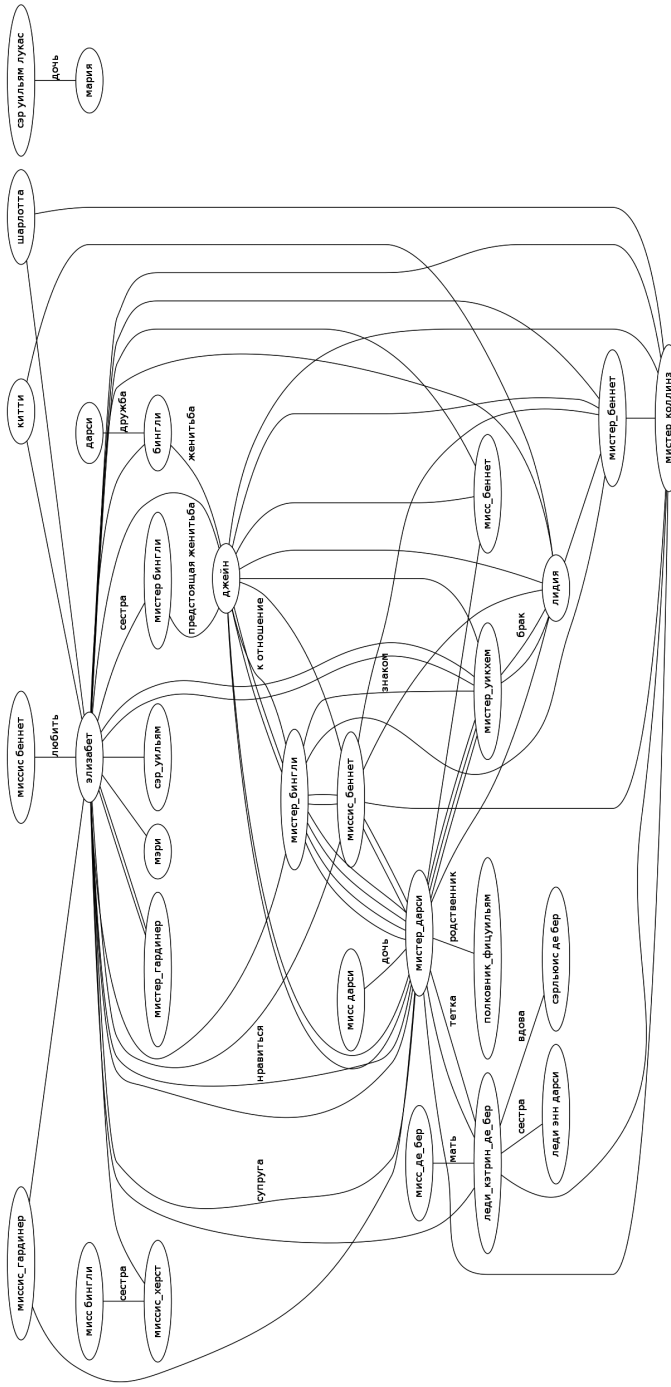
In this paper we presented a method of lexico-syntactic patterns for extracting social networks from fiction. This allowed us to take a systematic look at a large corpus of texts.

Our results thus far suggest further review of that method for more insights into social networks found in fiction.

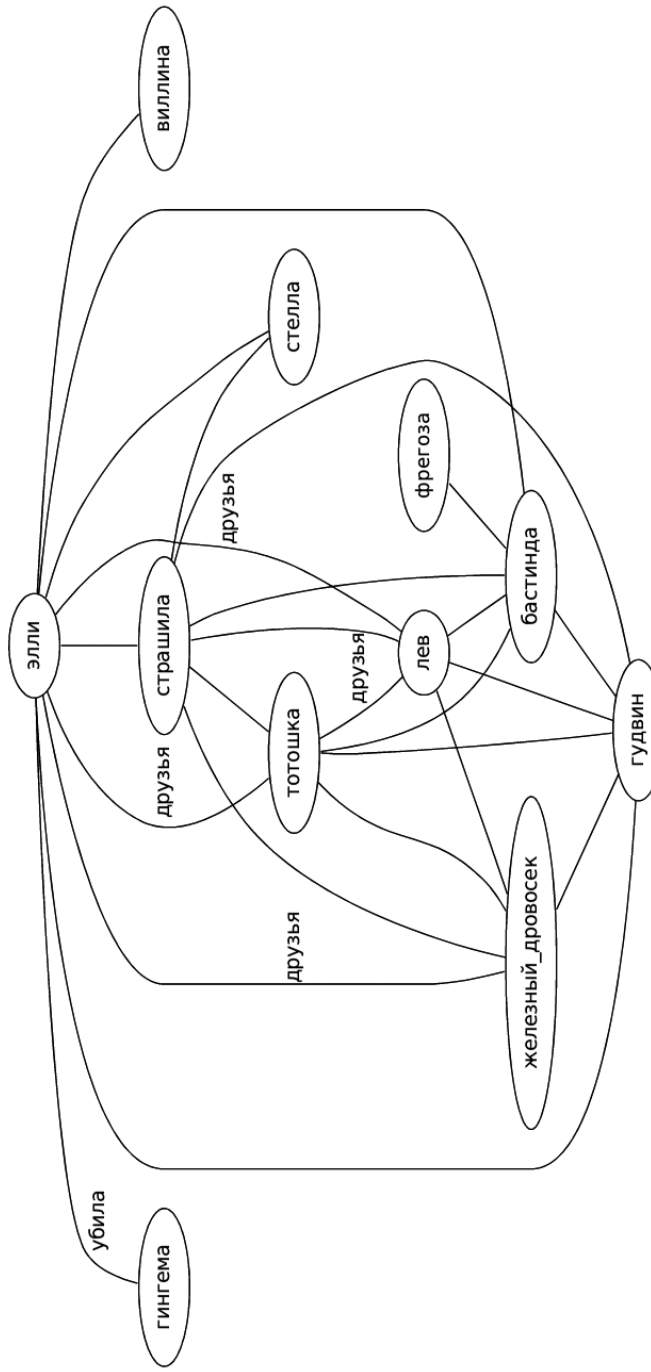
References

1. *Agarwal A.* (2011). Social Network Extraction from Texts: A Thesis Proposal. ACL2011.
2. *Elson D., Dames N. and McKeown K.* (2010), Extracting Social Networks from Literary Fiction, ACL2010, Uppsala, Sweden.
3. *He H., Barbosa D., Kondrak G.* (2013). Identification of Speakers in Novels. ACL2013
4. *Hearst M. A.* (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. The Fourteenth International Conference on Computational Linguistics, Nantes, France.
5. *Jockers M.* (2013). Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities). University of Illinois Press; 1st Edition.
6. *Lee J. and Yeung C. Y.* (2012). Extracting Networks of People and Places from Literary Texts. ACL2012.
7. *Makazhanov A., Barbosa D., Kondrak G.* (2012). Extracting Family Relations from Literary Fiction. Unpublished manuscript.
8. *Marazzato R., Sparavigna A. C.* (2014). Extracting Social Networks of Characters and Places from Written Works with CHAPLIN. arXiv.org, Cornell University Library.
9. *Michel J-B., Liberman A.* (2010). Quantitive Analysis of Culture Using Millions of Digitized Books. Science journal.
10. *Moretti F.* (2007). Graphs, Maps, Trees: Abstract Models for a Literary History Verso, London.
11. *Moretti F.* (2011). Network Theory, Plot Analysis. Pamphlet of Stanford Literary Lab.
12. *Park G.-M., Kim S.-H., Hwang H.-R., Cho H.-G.* (2013). Complex System Analysis of Social Networks Extracted from Literary Fictions. International Journal of Machine Learning and Computing.
13. *Sugiyama K., Yoshikawa M., Hatano K., Uemura S.* (2003). Extracting Information on Protein-Protein Interactions from Biological Literature Based on Machine Learning Approaches. Genome Informatics.

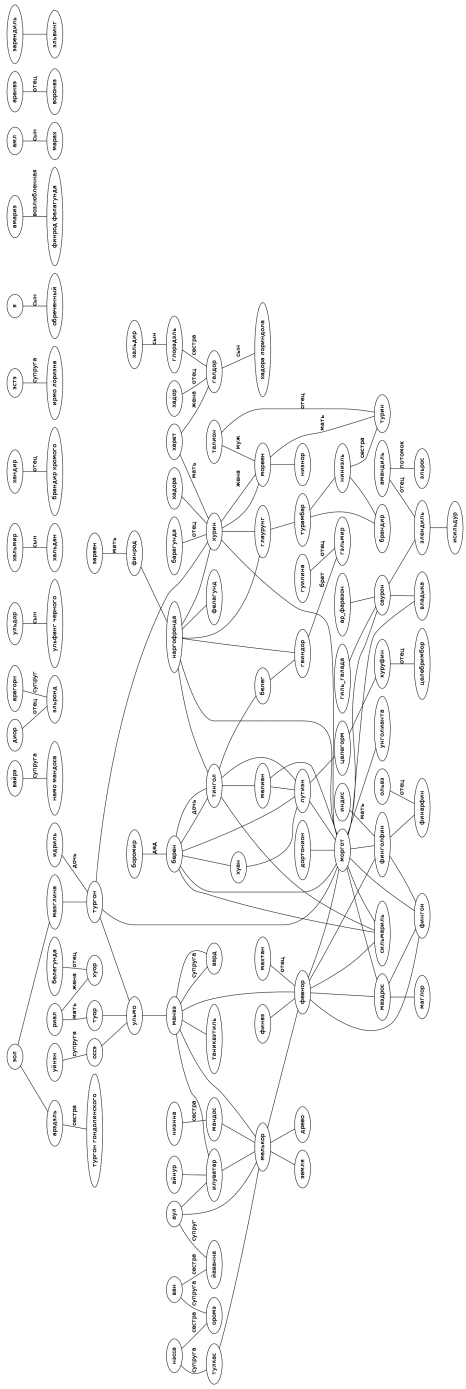
Appendix 1. J. Austen “Pride and Prejudice”



Appendix 2. A. Volkov “Wizard from Emerald City”



Appendix 3. J. Tolkien “Silmarillion”



Appendix 4. N. Kun “Legends and Myths of Ancient Greece”