Tolstoy Digital: Mining Biographical Data in Literary Heritage Editions

Anastasia Bonch-Osmolovskaya, Matvey Kolbasov
National Research University Higher School of Economics
20 Myasnitskaya Ulitsa; Moscow, Russia; 101000
abonch@gmail.com, matveykolbasov@yandex.ru

Abstract

This paper presents a solution for mining the biographical information from commentaries on Leo Tolstoy's letters. It is implemented as a part of Tolstoy Digital Project – a semantically marked-up web publication of the 90-volume complete collection of Leo Tolstoy's works. Extraction of relevant biographical information will be used to create an open database for all the persons who were somehow connected with Tolstoy or Tolstoy's works. The paper also accounts for various subtleties of the commentary apparatus and pays special attention to specific difficulties of biographical information extraction, such as the problem of defining the boundaries of expressions denoting profession, or the problem of non-standardized syntactic constructions for kinship relations.

Keywords: Leo Tolstoy, commentary apparatus, biographical database, semantic edition

1. Project Description

The Tolstoy Digital project[1] aims to prepare a web-published semantically marked-up version of the 90-volume complete collection of Leo Tolstoy's works[2]. The digital version of the 90-volume edition has become easy to access thanks to the mass crowdsourcing campaign "All of Tolstoy in one click"[3] . The next step of Tolstoy digitization is devoted to the semantic tagging of Tolstoy's text and the creation of a comprehensive database of all of the additional reference information that goes along with the Tolstoy oeuvres and private archive. The 90-volume edition (Tolstoy 1928-1964) comprises an exhaustive critical apparatus, which contains relevant information on Tolstoy's works, life, and other people connected with him. Current research, done as part of the Tolstoy Digital project, presents an on-going work in fact-extraction from literary commentaries. The edition contains 21 volumes of letters, dated from 1844 till 1910, the year of Tolstoy's death. Each letter is followed by a detailed commentary, where a biographical reference to the addressees and persons mentioned in the letter is provided. Our aim is to analyze unstructured text of a commentary, to extract person names and relevant biographical information, and to use TEI semantic annotations for relevant mark-ups (Barnard, et al 1995). Extracted data will be aggregated and stored in a reference database. The database will be linked with the text of the semantic edition.

The main aim of our current experiment was to estimate the effectiveness of a rule-based approach that could be used for data analysis and fact extraction for our type of texts. We started with the most basic concepts and facts: names, dates, professions and family relations. The problem of automatic named entity recognition and fact extraction seems to be very well elaborated, see (Grishman 2003, and Jurafsky Martin 2009) for a comprehensive outline and reference list. Still, information extraction in academic philological texts pursues slightly different aims than mining the news flow: there are graphical, linguistic and conceptual problems. Specific graphics (such as abbreviations, inner references, font highlighting) may have non-trivial semiotic functions. Some patterns that are uncommon in general language may be used to render the latent recurring logics of the commentary structure. Finally, the notions that are relevant for characterizing historical events or persons may have little in common with the conceptual space of data mining in the web (the domain which is primarily regarded in all the works on information extraction). This is the reason why the task of biographical information extraction from academic editions cannot be solved with the help of already existing solutions, but requires some specific process modifications. This paper aims to report on the first steps that we have made in this direction. We used a rule-based toolbox Tomita to write and apply some basic grammar rules that are used to extract relevant ontology concepts. In part (2) we will present a short outline of the biographical ontology we are going to elaborate and we will briefly describe preliminary data preparation. Part (3) will be dedicated to the description of our grammar and the rules for biographical fact extraction. Evaluation and analysis are presented in part (4).

2. Textual material and basic ontology

The commentary apparatus of epistolary volumes of the complete 90-volume edition consists of letter commentaries, which constitute about 40% of entire text. Usually they are organized in a non-structured way as a sequence of factual comments. Some of them are hard to classify, and many of them seem not to fit to any database, making them redundant. This lack of explicit structure can be explained by the fact that the commentaries have been created by different authors, and

[1] See the project description here: http://tolstoy.ru/projects/tolstoy-digital/

[2] http://tolstoy.ru/creativity/90-volume-collection-of-the-works/

[3] See, for example, review in the Guardian: http://www.theguardian.com/books/2013/oct/16/all-leo-tolstoy-one-click-project-digitisation

each of them had used his own text template. As a result, commentaries are represented as an accompanying text, but not as an enumeration of properties and parameters of a common database. An example of a commentary is provided in Table 1.

Russian	English
Анатолий Иванович Фаресов (1852—1928), публицист-народник. [Судился по «делу 193»; был амнистирован в 1880 г. и перешел в лагерь умеренных либералов.] Сотрудничал в «Новостях», в «Неделе» и других изданиях. Познакомился с Толстым 1 февраля 1898 г. [и оставил неопубликованные воспоминания о нем]. Статья Фаресова под заглавием «Ко вчерашнему происшествию редакции в «Недели» была напечатана в «Новом времени», № 7208 от 23 марта	Anatoly Ivanovich Faresov (1852-1928), writer-populist. [He was tried for "Case 193"; He was pardoned in 1880 and joined the camp of moderate liberals.] He worked in the "News," in the "Week," and other publications. He had met with Tolstoy February 1, 1898 [and left unpublished memories of him]. Faresov's article titled By yesterday incident edition of "Week"" was published in the "New Times" , № 7208 of 23 March

Table 1. An example of biographical commentary in the 90-volume Tolstoy's edition.

This example illustrates the complexity of the information in the comments. We have marked with square brackets all information that cannot be structurally formalized. That means that though we have at our disposable a considerable corpus of short biographical texts, but it is still a collection of unstructured texts with high degree of lexical variation and a mixture of relevant and excessive factual statements. We aim to transfer an unstructured commentary text into a structured database with given semantic relationships between the elements. For the first stage of our project, we limit ourselves to a small list of relationships. Table 2 shows the relevant facts, which have been filtered out of the text, given in table 1.

Text	Type of the information	Current relevance of the information
Anatoly Ivanovich Farezov (1852-1928)	The head of biographical information	YES
writer-populist	Main fact	YES
He was tried for "Case 193"; He was pardoned in 1880 and joined the camp of moderate liberals.	Additional political fact	NO
He worked in the "News," in the "Week," and other publications.	Additional profession fact	NO
He had met with Tolstoy February 1, 1898	Tolstoy fact	YES
and left unpublished memories of him.	Additional publicistic fact	NO
Faresov's Article titled "By yesterday incident edition of "Week"" was published in the "New Times" , № 7208 of 23 March	Faresov's publicistic fact	YES

Table 2. The choice of facts and relationships to be extracted from the commentaries.

The corpus of letters, used for rule development and testing, was based on the 63rd volume of the Complete edition, which contains 281 letters by Tolstoy of the period of 1880-1886. We have created a manual markup of 50 letters using the annotation module from GATE framework (Cunningham, et al 2002). We annotated all professions, kinship terms and relationships, and birth and death dates. The observations made upon this small annotated corpus were used to develop grammatical rules for biographical fact extraction. For example, we have found that the mention of a person, who has not been commented on before in the text, corresponds to a certain robust text structure. We called it a pattern of first mention. It consists of a personal name, dates of life, and additional information, which may be a profession or kinship references to other persons (first of all, to

Tolstoy). Finally we have built a small ontology, comprising the main entities and their attributes and matched in with our annotated sentences. The ontology is presented in Figure 1. The text spans referring to annotation are printed in bold.

<has name: first, middle, last>

RUS: ***Николай Николаевич Страхов*** *(1828—1896) — критик и философ.*

ENG: ***Nikolay Nikolayevich Strahov*** *(1828—1896) — critic and philosopher.*

<has maiden name>

RUS: *Анастасия Васильевна Дмоховская, урожд.* ***Воронец***. *О ней см. в т. 49*

ENG: *Anastasia Vasil'yevna Dmohovskaya, maiden name* ***Voronec***. *See vol.49.*

<has marriage name>

RUS: *Евдокия Александровна Новосильцева (р. 1861 г.), в замужестве* ***Регекампф***.

ENG: *Evdokiya Alexandrovna Novosil'ceva (born 1861), marriage name* ***Regekampf.***

<birth date>

RUS: *Владимир Иванович Даль (* ***1801*** *— 1872) — известный лексикограф и этнограф.*

ENG: *Vladimir Ivanovich Dal' (* ***1801*** *— 1872) — famous lexicographer and ethnographer.*

<death date>

RUS: *Михаил Александрович Энгельгардт (р. 1861 г. — ум.* ***21 июля 1915*** *г.)*

ENG: *Mihail Alexandrovich Engelgardt (born 1861 — die* ***21 July 1915****)*

<social status>

RUS**:** ***Гр.*** *Сергей Николаевич Толстой (1826—1904) — старший брат Льва Николаевича.*

ENG**:** ***Earl*** *Sergey Nikolayevich Tolstoy (1826—1904) — the elder brother of Leo Tolstoy.*

<social status:place>

RUS: *Иван Васильевич Сютаев (р. 1856), крестьянин* ***дер. Шевелино***

ENG: *Ivan Vasil'evich Sutaev (born 1856), peasant of* ***Shevelino village***

<profession>

RUS: *Афанасий Афанасьевич Фет (Шеншин) (1820—1892) —* ***поэт.***

ENG*:* *Afanasiy Afanas'evich Fet (Shenshin) (1820—1892) —* ***poet.***

<profession:date>

RUS: *Михаил Матвеевич Стасюлевич (1826—1911) — общественный деятель, историк и публицист,* ***с 1865 года*** *редактор-издатель журнала «Вестник Европы».*

ENG: *Mihail Matveevich Stasulevich (1826—1911) — social activist, historian and publicist,* ***since 1865*** *editor and publisher of the magazine "Herald of Europe".*

<kinship with: name>

<kinship type: son/daughter/father/mother/husband/wife/widow/widower>

RUS**:** ***Лев Львович*** *(р. в 1869 г.), третий* ***сын Толстого***.

ENG**:** ***Lev Lvovich*** *(b. 1869), the third* ***son*** *of* ***Tolstoy***.

< friend of: name>

<friendship type: acquaintance/friend/colleague>

RUS**:** ***Дмитрий Алексеевич Дьяков*** *(1823—1891) — сын Алексея Николаевича и Ирины Дмитриевны, рожд. Полторацкой,* ***друг Толстого***.

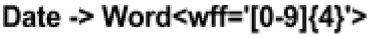
Figure 1. The ontology of biographical facts and attributes in commentaries to Tolstoy's letters.

For the first stage of our research, we decided to consider the most basic categories, such as name, dates of life, profession, and kinship. We have created several grammatical rules and have evaluated those rules with the help of our annotated commentary corpora.

3. Methods

We used Tomita parser to create rules for fact extraction (Tomita 1984, 1985). Tomita parser is a free NLP platform customized for creating small and light information extraction modules. It can be used as a tool for extracting structured data (facts) from texts by context-free grammars and dictionaries of keywords. The API for Tomita parser has been developed by the team of Yandex.ru[4] and is available for free download. Tomita provides modules of morphological analysis, as well as ready-made rules for extracting names and numbers. Tomita grammars consist of rules. The user may create his or her own grammars and dictionaries for a certain language. To construct a grammar, the user should write a number of transducer rules, for which one can use regular expressions, words lists, and other rules, built-in or composed. Each rule has a left and a right side; the transducer operation is denoted by arrow, separating left and right sides. The left side can be only represented by a terminal, while the right side can be both terminals or nonterminals. An example of a rule sequence can be found in Figures 2-3.

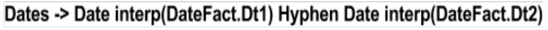
[4] https://tech.yandex.ru/tomita/



```
Date -> Word<wff='[0-9]{4}'>
```

Figure 2. Rule for date extraction.

This rule says that a date is a number, consisting of four characters from 0 to 9. Then, we can define date chains (for example dates of birth and death) using previously set rules. The operator *interp* in Figure 3 assigns the extracted text span to a specific fact. Here Dt1 stands for the birth date, while Dt2 corresponds to the date of death :



```
Dates -> Date interp(DateFact.Dt1) Hyphen Date interp(DateFact.Dt2)
```

Figure 3. Rule for date chains extraction.

Using Tomita parser we have written a grammar with specific rules and dictionaries aimed to extract the following biographic facts from letter commentaries: ProfessionFact (person's name, dates of life and professions) and FamilyFact (kinship). We quickly realized that the built-in date-extraction and name-extraction rules are not well-sutied to philological commentaries, so we also made some specific modifications. For description of kinship relations of a person, we have created several dictionaries containing different kinship types. The kinship dictionary is applied after the pattern is recognized. If no kinship term is detected in the pattern, then the rules that extract profession are applied. Overall there have been created 31 rules and 3 dictionaries. The processing scheme is presented by Figure 4.

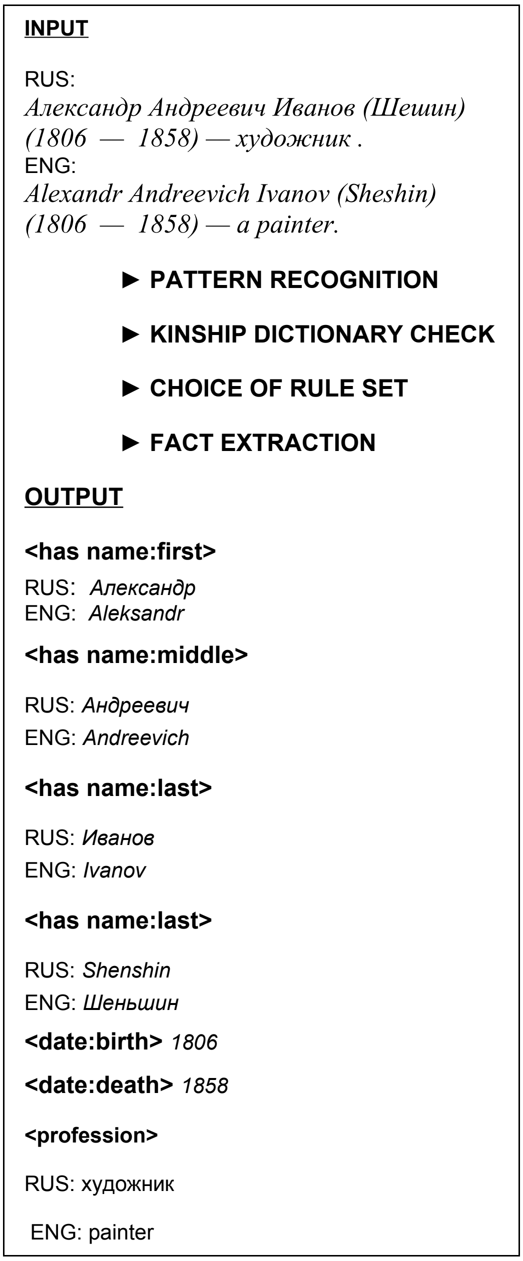
4. Results and analysis

The performance rank of these rules has been measured with the help of testing commentary corpus of 560,000 tokens. The results are considerably acceptable for precision, but are much lower for recall, which is quite typical for rule-based approaches. The overall score for both facts is demonstrated in Table 3

Fact name	Precision	Recall	F-measure
ProfessionFact	1	0,82	0,9
FamilyFact	0,78	0,3	0,43

Table 3. Evaluation of biographical fact extraction.

The analysis of results reveals that some important problems lie in the conceptual part. First of all, it is problematic to define a strict classification of what may (or may not) be considered as a profession. Thus in example 1 the profession extracted is a *peasant*. But in context of biographical commentaries this person's origin (*village Shevelevo*) becomes even more important than the profession of peasant itself. It's obvious that Tolstoy had met a lot of peasants during his life, so any special attributes are of much value in the context. There remains a problem of distinguishing those cases, which are meaningless without additional information.



INPUT

RUS:
Александр Андреевич Иванов (Шешин)
(1806 — 1858) — художник .
ENG:
Alexandr Andreevich Ivanov (Sheshin)
(1806 — 1858) — a painter.

► **PATTERN RECOGNITION**

► **KINSHIP DICTIONARY CHECK**

► **CHOICE OF RULE SET**

► **FACT EXTRACTION**

OUTPUT

<has name:first>
RUS: *Александр*
ENG: *Aleksandr*

<has name:middle>
RUS: *Андреевич*
ENG: *Andreevich*

<has name:last>
RUS: *Иванов*
ENG: *Ivanov*

<has name:last>
RUS: *Shenshin*
ENG: *Шеньшин*

<date:birth> *1806*

<date:death> *1858*

<profession>
RUS: художник
ENG: painter

Figure 4. Extracting information from text patterns.

Another problem is the boundaries of the text span which refers to the profession itself. Thus, in example 2, we see a description of Shamil's social activities. So the question is what part of the complicated NP should be extracted as Shamil's profession. Should it be the leader and consolidator or the full NP (*leader and **consolidator** of hillmen of Daghestan and Chechnya)*?

1) *Василий Кириллович Сютаев (1819—1892) — ... **крестьянин** дер. Шевелино.*

*Vasiliy Kirillovich Syutaev (1819—1892) — ... **peasant** of village Shevelino.*

2) *Шамиль (1797—1874) — знаменитый **вождь** и **объединитель** горцев Дагестана и Чечни в их борьбе с русскими.*

Shamil (1797-1874) – a popular ***leader*** *and* ***consolidator*** *of hillmen of Daghestan and Chechnya in their struggle with Russians.*

To resume, if such types of professions as writer, musician, and philosopher are good categorizers, that allow to define a group of people in one way or another connected with Tolstoy or his work, such status as a peasant or consolidator are meaningless without their attributes (genitive groups), in this case – *"village Shevelino"* and "*hillmen of Daghestan and Chechnya in their struggle with Russians."* Accordingly, the significant problem is with the boundary of the nominal group that determines the professional status. The second problem is concerned with intricate chains of kinships that have a very specific syntax, as shown in Example 3. In this phrase there is an inversion (see Figure 5), probably made especially for publication, so that Sophia Tolstaya, wife of Leo Tolstoy, would take a more prominent position.

3) *Александр Михайлович Кузминский ...* ***муж сестры Софьи Андреевны, Татьяны Андреевны*** *(1846—1925).*

Alexandr Mihaylovich Kuzminskiy ... ***the husband of the sister of Sofiya Andreevna, Tatiana Andreevna*** *(1846—1925).*

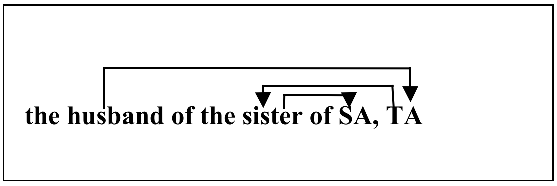


Figure 5. Complicated syntactic structures in kinship patterns.

Another problem is the discrepancy between singular and plural forms when it comes to the descriptions of relations between a person and a family group, as shown in Example 4.

4) *Николай Михайлович Нагорнов (1845—1896) — сын* ***Михаила Михайловича и Надежды Ивановны Нагорновых.***

Nikolay Mihaylovich Nagornov (1845—1896) — the son of ***Mihail Mihaylovich and Nadejda Ivanovna Nagornovs.***

Unlike profession extraction which shows good quality by rule-based approach, kinship relation patterns seem to be much less regular, so perhaps it is worth trying to extract them using algorithms of machine learning. In general, during the next stage of the research we intend to develop the ontology (i.e. add important locations and relations between person and location), to process all the 31 volumes of letters, to make an open database with all the persons connected to Tolstoy, and to provide every input referring to a person with a short semantically marked-up biography. The database is to be used as an interlinked reference base to Leo Tolstoy's 90 volume edition, and also as an aggregator of exterior information from other sources. The specific syntactic constructions, that are intrinsic for the commentary genre (such as abundance of names and titles in a sentence, special means of reference inside text, etc.) will be certainly taken into account.

References

Barnard, David T., et al. "Hierarchical encoding of text: Technical problems and SGML solutions." Text Encoding Initiative. Springer Netherlands, 1995. 211-231.Computers and the Humanities 29.3 (1995): 211-231.

Cunningham, Hamish, et al. "GATE: an architecture for development of robust HLT applications." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

Grishman, Ralph. "Information extraction: Techniques and challenges." Information extraction a multidisciplinary approach to an emerging information technology. Springer Berlin Heidelberg, 1997. 10-27.

Tolstoy, Lev. Polnoe sobranie sochinenij v 90 tomah, Moskva, 1928-1954. (Tolstoy, Leo. Complete collected works in 90 volumes, Moscow 1928-1954)

Tomita, Masaru. LR parsers for natural languages. COLING. 10th International Conference on Computational Linguistics. 1984. P. 354—357.

Tomita, Masaru. An efficient context-free parsing algorithm for natural languages. IJCAI. International Joint Conference on Artificial Intelligence. 1985. P. 756—764.

Acknowledgements

This study (research grant No 15-06-99523 A) was supported by the Russian Foundation for Basic Research in 2014-2015.