Automatic Compilation of Person's Information Portraits as an Instrument of Historical Research

Anna Glazkova[1], Valery Kruzhinov[2], and Zinaida Sokova[2]

[1] Tyumen State University, Institute of Mathematics and Computer Science,
Tyumen, Russia
anna_glazkova@yahoo.com,
[2] Tyumen State University, Institute of History and Political Sciences, Tyumen,
Russia
{v.m.kruzhinov,z.n.sokova}@utmn.ru

Abstract. The authors propose an approach to the compilation of person's information portraits. The development of this approach is relevant for the discovery of information useful for historical research. The article proposes a methodology for creating information portraits consisting of two stages: named-entity recognition and fact extraction. The authors carry out an experimental verification of the described approach. The proposed approach is implemented within the prototype of the information system. The implemented technology is intended for a wide range of biographical researchers and persons interested in history.

Keywords: data extraction, information portrait, historical methodology, natural language processing, parsing.

1 Introduction

The historical research, generally, is related to the study of biographical facts. These facts are diverse. For example, the historian is interested in social, economic, political and cultural aspects of person's lives. The search for the listed types of information is possible with the help of various sources. In particular the Internet and electronic resources are actively used in modern practical historical science.

The search for diverse text information on the Internet is associated with some difficulties. For example, scientists can not always clearly formulate a search query. Some useful facts often seem indirect in the text. Therefore, the historian is forced to read a huge amount of information to obtain these facts and view a large number of web pages that can contain interesting biographical information.

The idea of the technology presented in this article is to automate the biographical facts extraction. The peculiarity of this project is its principled focus on the needs of historians and biographers. We are developing a technology that allows us to extract named entities and the corresponding facts from the text. The extracted facts about the person should be stored in the database, then other researchers will also have access to them.

In this paper we use the term «information portrait». An information portrait of a person is a named entity that denotes a historical personality and a set of corresponding biographical facts.

Related Work

Compilation of person's information portraits implies the solution of the following tasks:

- named-entity recognition;
- fact extraction from texts.

Although named-entity recognition (NER) is a widely studied problem, its solution is highly dependent on the subject area. NER has been applied to different types of text: email [1], Wikipedia articles [2], tweets [3,4], news articles [5], etc. Named entities extraction from historical and biographical texts was carried out in the works of K. Byrne [6], C. Grover et al.[7], T. Packer et al.[8]. B. Batjargal et al. [9] proposed the approach to NER from traditional texts. A detailed comparison of NER tools applied to biographical texts is given in the article [10]. The presented instruments showed quite accurate results, but their using in their pure form is impossible for Russian texts due to the specific features of the Russian anthroponymic. Existing developments for the Russian language (for example, [11]) need to be refined for historical texts.

The issues of extracting biographical facts from the text were considered in the article of I. M. Adamovich and O. I Volkov [12]. The authors describe a technology that represents the fact as a tree structure. The root of the constructed tree is a fact (e.g. «birth»), the related entities are stored in leaves. Also historical facts extraction were discussed in the article of A. Cybulska and P. Vossen [13]. The authors applied to fact extraction a generic fact mining system KYOTO. D. Hienert and F. Luciano [14] proposed two approaches to extract events from Wikipedia: events extracting from the main article text and the creation of events from the article itself.

2 Methods

In this work, the compilation of person's information portraits is performed in 2 steps.

Step 1. Search for named entities that label people.
Step 2. Extracting facts related to the named entity.

Person's Names Recognition

In this study, we adopted the following rules for identifying named entities labelling persons.

First of all, we search for named entities that already presented in the database. Then the words of the text are checked by dictionary of personal names and surnames of the Russian language. This dictionary was compiled manually based on publicly available data on the Internet.

The word is supposed to be a named entity or a part of a named entity if it is written with a capital letter. In the case when after the word which is written with a capital letter there are other words that begin with a capital letter, they are also included in the supposed named entity. If there are punctuation marks between the words, these words are marked as different named entities. When there are specific words between the words beginning with a capital letter (for example, «de» or «de la»), specific words are ignored, the search for words beginning with the capital continues.

There are a number of difficult cases, for them we were forced to create separate templates. For example, word combinations like «prince Galitsy Vladimir» or «president of RZhD Belozerov» are analysed using the vocabulary of posts and professions and taking into account the case endings. As a result, the parser retrieves the named entities «Vladimir» and «Belozerov» and respectively the relevant facts «prince Galitsky» and «president of RZhD». Another difficult case is constructions like «son of Peter Stolypin Arkady Stolypin». In this example, the parser extracts a sequence of 4 words.

In most cases, the parser distinguishes constructions with homogeneous parts of the sentences such as «Petya and Nikolay Rostov» from constructions like «prince Andrei and Nikolai Rostov» with the help of the Russian case forms of surnames (for this examples the forms are «Rostovy» and «Rostov» in the transliteration from the Russian). So, the result of the parser's work is a pair of entities «Petya Rostov» and «Nikolai Rostov» for the first case and «Andrei» and «Nikolay Rostov» for the second case.

The entity is not included in the list of personal names in the event that there is one of the following excluding attributes:

- quotes;
- one of the parts of the essence is an abbreviation;
- word-markers within, before or after the entity (in particular the words «ministry» or «university»).

At the next stage supposed named entities are checked for the presence of the following attributes:

- specific surname suffixes (-enco, -shvili, etc.) or patronymic (-vich, -vna);
- specific surname prefixes (Mac-, O'-, etc.);
- initials (one or two capital letters with dots, possibly a hyphen between the initials), after or before the initials there is a word with a capital letter (possibly two words associated with a hyphen);
- previous words «prince», «minister», etc.;
- verified pattern matching (examples of patterns are «Surname + Name + Patronymic» or «Name + Regnal number»).

If any of these attributes is present, the entity is marked as the person's name. Entities that do not have name attributes and excluding attributes are offered to the user for manual markup.

The flowchart for person's names recognition is shown in Fig. 1.

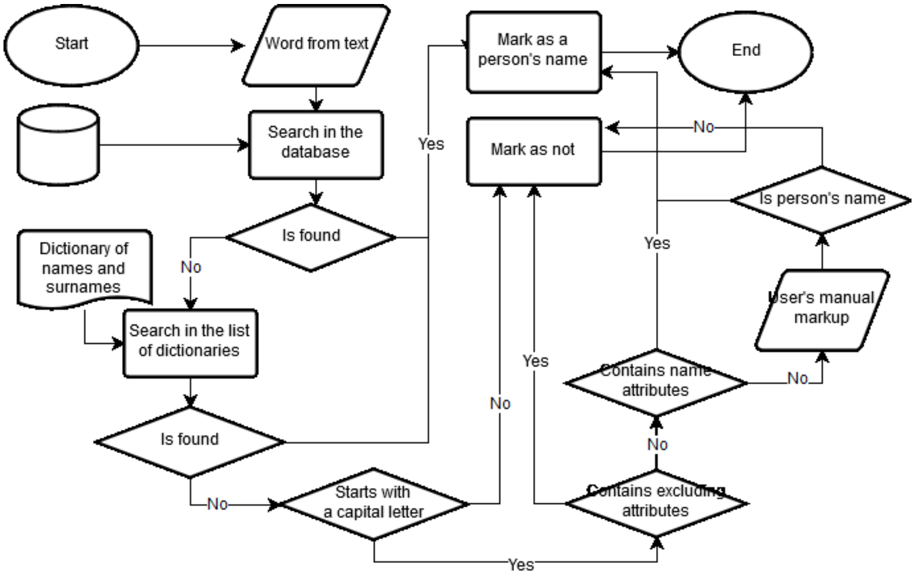

Fig. 1. Person's names recognition

Fact Extraction

For fact extraction the text is divided into sentences, complex sentences are splitted into simple sentences.

For the sentence where the named entity is found the following actions are performed:

- marking the words that are included in the sentence text (part of speech definition, case of nouns definition);
- fact extraction based on the templates.

Further, we establish the presence of a co-reference with the following sentences or with the following parts of a complex sentence. The co-reference is a binding of several different references in the text to one real object. It was experimentally determined that the co-reference occurs, as a rule, at a distance of not more than two sentences. If the connection is established, the following sentence also extracts the fact.

For example, given the text (translation from Russian): "In 1943 writer Alexander Solzhenitsyn was promoted to senior lieutenant. Next year he began to command a battery in an artillery brigade. On 8 July 1944 he was awarded the Order of the Red Star".

First of all, the text is divided into sentences. We get the following set of sentences:

1. In 1943 writer Alexander Solzhenitsyn was promoted to senior lieutenant.
2. Next year he began to command a battery in an artillery brigade.
3. On 8 July 1944 the writer was awarded the Order of the Red Star.

In the first sentence the parser finds a person's name. The first sentence is subjected to syntactic markup. Based on the results of markup, the text is transformed into the following facts:

- Alexander Solzhenitsyn <person>
- writer <fact> – null <date>
- was promoted to senior lieutenant <fact> – 1943 <date>.

The word «writer» in this example can be used as a synonym for the name «Alexander Solzhenitsyn».

In our parser we tried to resolve the co-referencing. One of the types of co-referencing is the anaphora, or the reference to the object by means of special pointers (for instance, pronouns). The second kind of co-referencing is synonymy. Both types are present in the text example.

An anaphoric relationship between the sentences 1 and 2 is established due to the pronoun «he» in the second sentence. So, the parser replaces the pronoun by the person's name and also extracts the fact. The parser converts the phrase «next year» into a date based on the last found year designation in the text. Similarly, an co-referencing relationship is found with the following sentence. Thus, the following information portrait is extracted from the text:

- Alexander Solzhenitsyn <person>
- writer <fact> – null <date>
- was promoted to senior lieutenant <fact> – 1943 <date>
- began to command a battery in an artillery brigade <fact> – 1944 <date>
- was awarded the Order of the Red Star <fact> – 08.07.1944 <date>

The flowchart for fact extraction is shown in Fig. 2.

3 Results

We conducted an experiment to test the proposed approach. A part of the experiment devoted to named-entity recognition were based on the open Russian text collection «Persons-1000» [15,16]. To compile information portraits we manually collected the corpus of 30 texts.

The result of the experiment was the accuracy of extraction. Accuracy was defined as the ratio of the number of extracted entities to the total number of entities.

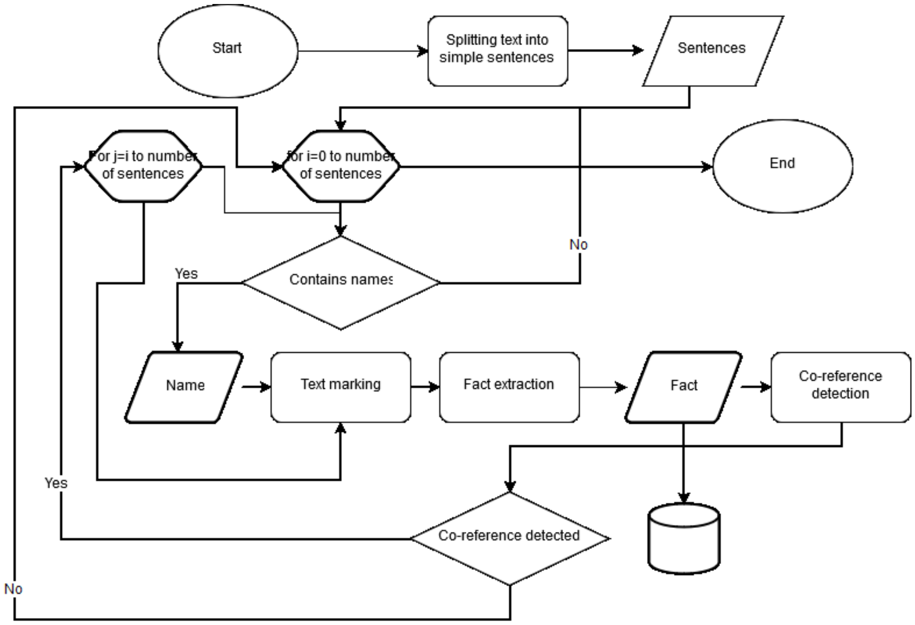

Fig. 2. Fact extraction

Based on the results of the experiments, the accuracy of person's names extraction was 91,73%. The accuracy of information portraits compiling was 70,96%. The study took into account only fully correctly compiled information portraits.

4 Discussion

In this paper we showed that person's information portraits can successfully be extracted from text, based on constructional clues and semantic type specification.

In future we will make an attempt to increase the accuracy of the parser. For this purpose, we plan to plan to establish links between named entities, improve the consideration of syntactic links and add more attributes to both name attributes and excluding attributes.

In the next stage of the project we will develop a comprehensive information system for the compiling of information portraits of historical personalities. This system will be in demand both by biographical researchers and by ordinary users interested in genealogy or searching for information about famous people.

Acknowledgements The authors wish to thank all those who participated in the field exercise and helped to make it a successful endeavour.

References

1. *Minkov, E., Wang, R. C., Cohen, W.* Extracting personal names from email: applying named entity recognition to informal text / in «HLT/EMNLP». P. 443–450. — Vancouver: ACL, 2005.
2. *Mohamed, M., Oussalah, M.* Identifying and Extracting Named Entities from Wikipedia Database Using Entity Infoboxes // International Journal of Advanced Computer Science and Applications. 2014. Vol. 7, N 5. P. 164–169.
3. *Derczynski, L., Maynard, D., Rizzo, D., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.* Analysis of named entity recognition and linking for tweets // Information Processing and Management. 2015. N 51. P. 42–49.
4. *Espinosa, K. H., Batista-Navarro, R., Ananiadou, S.* Learning to recognise named entities in tweets by exploiting weakly labelled data / in «Proceedings of the 2nd Workshop on Noisy User-generated Text». P. 153–163. — Osaka, 2016.
5. *Kanana, T., Kanaab, R., Al-Dabbas, O., Kanaab, G., Al-Dahoud, A., Fox, E. A.* Extracting Named Entities Using Named Entity Recognizer for Arabic News Articles // International Journal of Advanced Studies in Computers, Science and Engineering. 2016. Vol. 5. N 11. P. 78–84.
6. *Byrne, K.* Nested Named Entity Recognition in Historical Archive Text / in «Proceedings of ICSC-2007». P. 589–596. — Irvine, 2007.
7. *Grover, C., Givon, S., Tobin, R., Ball, J.* Named Entity Recognition for Digitised Historical Texts / in «Proceedings of the International Conference on Language Resources and Evaluation». — Marrakesh, 2008.
8. *Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K.* Extracting Person Names from Diverse and Noisy OCR Text / in «Proc. of the 4th workshop on Analytics for noisy unstructured text data». P. 19–26. — Toronto, 2010.
9. *Batjargal, B., Khaltarkhuu, G., Kimura, F., Maeda, A.* An approach to named entity extraction from historical documents in traditional mongolian script / in «IEEE/ACM Joint Conference on Digital Libraries ». — 2014.
10. *Atdag, S., Labatut, V.* A Comparison of Named Entity Recognition Tools Applied to Biographical Texts // in «2nd International Conference on Systems and Computer Science». P. 228–233. — 2013.
11. *Kuznetsov, I. P., Kozerenko, E. B., Charnine, M. M., Matskevich, A. .G., Nikolayev, V. .G., Somin, N. V.* Intelligent Systems for Entities Extractions Based on Extended Semantic Networks / in «Open Semantic Technologies for Intelligent Systems». P. 373–380. — Minsk: BGUIR, 2012.
12. *Adamovich, I. M., Volkov, O. I.* The system of facts extraction from historical texts // Systems and Means of Informatics. 2015. Vol. 25, N 3. P. 235–250.
13. *Cybulska, A., Vossen, P.* Historical Event Extraction From Text / in «Proceedings of 5th ACL-HLT Workshop on Language Technology on Cultural Heritage, Social Sciences and Humanities.» P. 39–43. — Portland: Association for Computational Linguistic, 2011.
14. *Hienert D., Luciano F.* Extraction of Historical Events from Wikipedia / in «ESWC 2012 Satellite Events.» P. 16–28. — Berlin: Springer, 2015.
15. *Vlasova N.A., Sulejmanova E.A., Trofimov I.V.* Message about the Russian-language collection for the task of extracting personal names from texts / in «Proceedings of the conference on computer and cognitive linguistics TEL'2014 «Language semantics: models and technologies» P. 36–40. — Kazan, 2014.
16. *«Person-1000»* Collection. URL: http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000. Date of access: 17.03.2017.