ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 8, August 2014

Towards Facts Extraction from Texts in the Polish Language

Tomasz Boiński, Adam Brzeski

Faculty of Electronics, Telecommunication and Informatics, Gdansk University of Technology, Poland

ABSTRACT: The Polish language differs from English in many ways. It has more complicated conjugation and declination. Because of that automatic facts extraction from texts is difficult. In this paper we present basic differences between those languages. The paper presents an algorithm for extraction of facts from articles from Polish Wikipedia. The algorithm is based on 7 proposed facts schemes that are searched for in the analyzed text. The analysis includes morphosyntactic tagging, named entity extraction and relation identification. The results acquired for an exemplary Wikipedia text is presented. We indicate the free word formation principle as the main difficulty in the Polish texts analysis. At the same time satisfactory performance of the tagging and analysis tools for the Polish language was confirmed in the conducted experiment.

KEYWORDS: natural language processing, text analysis, knowledge extraction, unstructured information, tagging, named-entity recognition

I. INTRODUCTION

Internet contains a lot of knowledge. It estimated that currently there are over 3.3 billion web pages [1]. Most of those pages are documents formed in natural language thus information (or facts) extraction from such documents was in the interest of researchers from the beginning of the Internet era.

So what is an information extraction? It is a process of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. In most of the cases this activity concerns processing human language texts by means of natural language processing (NLP).

The aim of this paper is to make a step towards the full automatic facts extraction in the Polish texts. Many researchers focus on the most widely used English language, which therefore has many tools available. Unfortunately those solutions, even those highly viable, do not perform well when used in conjunction with other natural languages. The nature of the Polish language makes it hard to apply the same rules as can be used for the English language. In this paper we focus on some basic ideas and problems that arise during our preliminary tests.

The structure of this paper is as follows. In the next chapter we present related work from the literature. Then the difference between Polish and English languages is presented. Next, the proposed way of extracting facts is described. Finally, we present the obtained results and the conclusions.

II. RELATED WORK

The problem of automated facts extraction plays a more and more important role in web pages processing. Especially if the information (or even knowledge) is contained within unstructured, natural language formatted texts.

For many years researchers and companies tried to tackle this problem using different approaches. The basic approach involved creation of patterns. In most cases those patterns were created before analysis of the text and then applied to search for matching facts [2]. Such approach introduced the need for proper complex patterns before the analysis of the texts and required supervision in the extraction process. Some approaches tried to eliminate this problem by providing means for automatic or semiautomatic pattern learning [3]. Finally modern approaches provide ontology based solutions eliminating the need for pattern creation and recognition [4], [5], [6], [7].

The problem of the polish language

The Polish language differs from English in many ways. The most important differences are:

1. The Polish language is formation free. Unfortunately the dominance of analytic languages, such as English or Chinese, makes research focuses primarily on the languages with fixed formation, while the language with free formation are less explored [8]. Tools like Świgra [9] or TaKIPI [10] solve this problem to some extent.
2. More complicated regular conjugation – Polish has more conjugation templates whereas English has fewer templates with far greater number of exceptions. Furthermore, the Polish language is further complicated by inflection [11].
3. Complicated declination – modern English, similarly as with conjugation, has very simple declination compared to Old English or Polish, however it has much more exceptions.
4. Combining declination with free word order makes sentences in Polish much more ambiguous than in English.

Until recently, the Polish language also lacked proper tools for automating common tasks like tagging, finding lemma of the word or named entities look-up. The situation changed with the development of Morfeusz [9], [12] which performs a morphological analysis for Polish sentences. Morfeusz became a base for Świgra, TaKIPI and recently Pantera [13]. All those tools are efficient taggers of the Polish language. Another useful application is Spejd [14], [15], [16], a tool for partial parsing and rule-based morphosyntactic disambiguation. Nerf [17] in turn allows extraction of named entities. Most of those tools require an extensive corpora, especially during the process of named entities extraction or coreference analysis. Such a common corpus were developed during recent years – The National Corpus of Polish [18], [19]. All those tools allow complex analysis of text in Polish, laying foundations for analysis and extraction of knowledge contained within documents formulated in natural language.

III. EXTRACTING FACTS FROM WIKIPEDIA

In our research we focused on extracting knowledge from Wikipedia articles. The main body of a Wikipedia article is rather loosely formatted with arbitrary chosen sections and text blocks. Also the content of each page is a natural language text without a formal structure. We attempted to extract the facts in a form of <subject, predicate, object> triples.

The test were done using Multiservice web site [20]. The general idea is to tag texts using Pantera and extract named entities using Nerf. Verbs (subst) and named entities (ne) of type [17]: persName, placeName, orgName and geogName are than mapped to subjects and objects, named entities of type date to objects. Pseudo participles (praet), participles (ppas) and prepositions (prep) are always treated as predicates. Adjectives (adj) were detected but ignored. The tags were taken from IPI PAN corpus tag syntax [21].

Two additional relations were introduced: isA and of. The isA relation introduces subsumption and can take named entities and verbs as subjects and verbs as objects. The of relation means that subject is related to object by some action, e.g. a boss is a chief of the company but the company is not subsumed by the boss. This relation can take verbs as subjects and verbs and named entities as objects.

The assignment of verbs and named entities to subjects and objects in triples depends on the morphosyntactic context it is used in. Currently we recognize the following schemes (square brackets ("[" and "]") means optional occurrence, pipe ("|") means alternative):

1. ppas date [prep ne]
2. ppas prep placeName
3. prep date [prep] (subst | [praet [subst [subst] [ne]]])
4. prep subst ne subst

5. subst prep ne subst [adj] [prep subst ne]
6. subst subst ([adj] | [prep subst [ne]])
7. subst ne

When a phrase matching one of the schemes is found, the words are connected with the main subject. Unfortunately currently the user has to select one of the verbs or named entities as the subject of the sentence.

IV. THE RESULTS

Preliminary research yielded some satisfactory results. Most of the facts were extracted. For example for Polish text "Bronisław Maria Komorowski (urodzony 4 czerwca 1952 w Obornikach Śląskich) – polski polityk, z wykształcenia historyk. Od 6 sierpnia 2010 prezydent Rzeczypospolitej Polskiej." ("Bronislaw Maria Komorowski (born June 4, 1952 in Oborniki śląskie) - Polish politician, educated as a historian. Since August 6, 2010 President of the Polish Republic.") [22] we acquire the following facts:

- Declaration: Bronisław Maria Komorowski (declaration of the main subject),
- urodzić w Oborniki śląski (born in Oborniki Śląskie),
- Bronisław Maria Komorowski urodzić 4 czerwiec 1952 (born June 4, 1952),
- Bronisław Maria Komorowski isA polityk (Bronisław Maria Komorowski isA politician),
- Bronisław Maria Komorowski isA historyka (Bronisław Maria Komorowski isA educated as historian),
- Bronisław Maria Komorowski isA prezydent (Bronisław Maria Komorowski isA president),
- prezydent od 6 sierpień 2010 (president since August 6, 2010),
- prezydent of rzeczpospolita polski (president of the Polish Republic).

As can be seen all of the facts were extracted correctly. Closer look however reveals some drawbacks of the existing tools. In the Polish language the basic form differs much from the one after declination. A reader familiar with Polish language can than find entities like "Oborniki śląski" or "rzeczpospolita polski" understandable but quite odd. The correct forms are "Oborniki Śląskie" and "Rzeczpospolita Polska" respectively. Unfortunately, in order to present the right form, the morphosyntax tagger would require a database of all named entities and their basic forms to properly formulate given named entity. The other problem with named entities are abbreviations. Usually "RP" stands for "Rzeczpospolita Polska". Our current solution will treat both of those entities as different ones. The same problem applies to normal verbs. The form of the education in the above example is incorrect. Instead of "historyka" it should be "historyk". This in turn is caused by difficulty of guessing the correct form of the base lemma (like singular or plural form, the proper declination of the original etc.). We plan to address those problems using Słowosieć (Polish version of WordNet) [23], [24], [25].

Further problems came with named entities consisting of more than one named entity. For example persName consists of at least one forename and surname. For our studies we decided to take into account the most complex form as one entity. In further studies we plan extracting additional information about each named entity based on its elements.

V. CONCLUSIONS

Much work has been done in the field of facts extraction from natural language texts. Recently at least 2 major research centers in Poland emerged that focus on automation of analysis of the Polish language. More and more tools are becoming available, leading towards the full analysis of the Polish language.

The biggest problem lies in the complexity of the Polish language. The multitude and complexity of conjugation and declination can be however solved by usage of proper morphological analyzers. Another issue lies in the freedom of formation. This highly complicates construction of templates that can be applied to the text.

With the constant development of supplementary tools and the experience gained through the research done for the English language, our preliminary research shows that a viable solution to facts extraction from documents formulated in the Polish language should be available soon.

REFERENCES

1. M. de Kunder, “World wide web size”, http://www.worldwidewebsize.com/, 2014, [Online: 18.07.2014].
2. R. Grishman, “Information extraction: Techniques and challenges”, in Information Extraction A Multidisciplinary Approach to an Emerging Information Technology. Springer, pp. 10–27, 1997.
3. N. Chambers and D. Jurafsky, “Template-based information extraction without the templates”, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, pp. 976–986, 2011.
4. B. Magnini, M. Negri, E. Pianta, L. Romano, M. Speranza, L. Serafini, C. Girardi, V. Bartalesi, and R. Sprugnoli, “From text to knowledge for the semantic web: the ontotext project” in SWAP, vol. 166, 2005.
5. D. C. Wimalasuriya and D. Dou, “Ontology-based information extraction: An introduction and a survey of current approaches”, Journal of Information Science, 2010.
6. J. Fan, A. Kalyanpur, D. Gondek, and D. A. Ferrucci, “Automatic knowledge extraction from documents”, IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 5–1, 2012.
7. H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt, “Automatic ontology-based knowledge extraction from web documents”, Intelligent Systems, IEEE, vol. 18, no. 1, pp. 14–21, 2003.
8. P. Skórzewski, “Wydajne algorytmy parsowania dla j˛ezyków o szyku swobodnym”, Ph.D. dissertation, Uniwersytet im. Adama Mickiewicza w Poznaniu Wydział Matematyki i Informatyki, 2014.
9. M. Woliński, “Komputerowa weryfikacja gramatyki Świdzińskiego”, Ph.D. dissertation, Instytut Podstaw Informatyki PAN, Warszawa, 2004.
10. M. Piasecki, “Polish tagger TaKIPI: Rule based construction and optimisation”, Task Quarterly, vol. 11, no. 1–2, pp. 151–167, 2007.
11. A. Przepiórkowski, “Slavonic information extraction and partial parsing”, in Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies. Association for Computational Linguistics, pp. 1–10, 2007.
12. M. Woliński, “Morfeusz—a practical tool for the morphological analysis of polish”, in Intelligent information processing and web mining. Springer, pp. 511–520, 2006.
13. S. Acedański, “A morphosyntactic brill tagger for inflectional languages”, in Advances in Natural Language Processing. Springer, pp. 3–14, 2010.
14. A. Przepiórkowski and A. Buczynski, “Shallow parsing and disambiguation engine”, in Proceedings of the 3rd Language & Technology Conference, pp. 340–344, 2007.
15. A. Buczyński and A. Wawer, “Shallow parsing in sentiment analysis of product reviews”, in Proceedings of the Partial Parsing workshop at LREC, pp. 14–18, 2008.
16. A. Buczyński and A. Przepiórkowski, “Spejd: A shallow processing and morphological disambiguation tool”, in Human Language Technology, Challenges of the Information Society, Springer, pp. 131–141, 2009.
17. A. Savary, J. Waszczuk, and A. Przepiórkowski, “Towards the annotation of named entities in the national corpus of polish”, in LREC, 2010.
18. A. Przepiórkowski, R. L. Górski, M. Lazinski, and P. Pezik, “Recent developments in the national corpus of polish”, in LREC, 2010.
19. A. Przepiórkowski, R. L. Górski, B. Lewandowska-Tomaszyk, and M. Lazinski, “Towards the national corpus of polish”, in LREC, 2008.
20. I. PAN, “Multiservice demo”, http://glass.ipipan.waw.pl/multiservice/, 2014, [Online: 09.08.2014].
21. M. Woliński, “System znaczników morfosyntaktycznych w korpusie IPI PAN”, Polonica, vol. 43, pp. 39–55, 2003.
22. Wikipedia, “Bronisław komorowski”, http://pl.wikipedia.org/wiki/Bronis%C5%82aw_Komorowski, 2014, [Online: 09.08.2014].
23. E. Rudnicka, M. Maziarz, M. Piasecki, and S. Szpakowicz, “Mapping plWordNet onto Princeton WordNet”, 2012.
24. M. Maziarz, M. Piasecki, and S. Szpakowicz, “Approaching plWordNet 2.0”, Matsue, Japan, 2012.
25. 26. M. Piasecki, S. Szpakowicz, and B. Broda, “A Wordnet from the Ground Up”, Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf, 2009, [Online: 09.08.2014].