UIMA-based Language Processing of Polish

Maciej Ogrodniczuk

Institute of Computer Science
Polish Academy of Sciences
ul. Ordona 21, Warsaw, Poland

maciej.ogrodniczuk@ipipan.waw.pl

Abstract

The ATLAS project, started in March 2010, intends to create the multilingual language processing framework integrating the common set of linguistic tools for a group of European languages, among them Polish. The chained tools are producing multi-level UIMA-encoded annotation of texts which can be used by higher-level Web applications for complex language-intensive operations such as automated categorization, information extraction, machine translation or summarization.

This paper concentrates on properties of language of processing chains integrated for Polish and findings specific to this integration. Inflectional characteristics of Polish offers the possibility to present a few more advanced functions such as multiword unit lemmatisation, vital for real-life presentation of extracted phrases.

1. Introduction

The ATLAS project (Belogay et al., 2011; Ogrodniczuk and Karagiozov, 2011; Karagiozov et al., 2011) offers the possibility to test integration of several NLP tools for Polish together with other project languages — Bulgarian, English, German, Greek and Romanian. To demonstrate capabilities of the framework, three linguistically-aware online services have been built on top of it: i-Publisher (Web-based content management platform), i-Librarian (a digital library of scientific works) and EUDocLib (site for browsing and searching through EUR-LEX documents).

2. Polish Language Tools

A number of existing tools for processing Polish were used by the project. According to the agreed annotation model, supporting the target application of linguistic data, particularly their display to the user, the tools were in certain cases reconfigured and extended to provide additional information such as ready-to-display normalized versions of base forms of identified multiword noun phrases.

2.1. Morfeusz and Pantera

Morfeusz (Woliński, 2006) is a morphological analyzer for Polish, also used for sentence- and token-level segmentation and lemmatisation of texts before the morphological part is applied. It uses positional tags starting with POS information followed by values of morphosyntactic categories corresponding to the given part of speech (Przepiórkowski and Woliński, 2003). Current version of the tool, Morfeusz SGJP, is based on linguistic data coming from The Grammatical Dictionary of Polish (Saloni et al., 2007).

Pantera (Acedański, 2010; Acedański and Gołuchowski, 2009) is a recently developed morphosyntactic rule-based Brill tagger of Polish. It uses an optimized version of Brill's algorithm adapted for specifics of inflectional languages. The tagging is performed in two steps, with a smaller set of morphosyntactic categories disambiguated in the first run (part of speech, case, person) and the remaining ones in the second run. Due to free word order nature of Polish the original set of rule templates as proposed by Brill has been extended to cover larger contexts.

2.2. Spejd and Multi-Word Expression Lemmatiser

Spejd (Przepiórkowski, 2008; Przepiórkowski and Buczyński, 2007) is an engine for shallow parsing using cascade grammars, able to co-operate with TaKIPI for tokenization, segmentation, lemmatisation and morphologic analysis. Parsing rules are defined using cascade regular grammars which match against orthographic forms or morphological interpretations of particular words. Spejd's specification language is used, which supports a variety of actions to perform on the matching fragments: accepting and rejecting morphological interpretations, agreement of entire tags or particular grammatical categories, grouping (syntactic and semantic head may be specified independently). Users may provide custom rules or may use one of the provided sample rule sets.

Spejd is the basis for the nominal groups lemmatiser (Degórski, 2011) developed throughout the project which combines the lemmatisation task with shallow parsing. The parsing structures are used as bases for lemmatisation schemata written separately for each grammar rule and operating on the strings and structure matched by that rule.

2.3. Named Entity Recognizer

Polish NER tool (Savary et al., 2010; Waszczuk et al., 2010) is a statistical CRF-based named entity recognizer trained over 1-million manually annotated subcorpus of the National Corpus of Polish (Przepiórkowski et al., 2008) and

The work reported here was carried out within the Applied Technology for Language-Aided CMS project co-funded by the European Commission under the Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467). The authors would like to thank all representatives of project partners for their contribution.

successfully used in the process of automated annotation of its total 1 billion segments.

The annotation scope is consistent with general ATLAS annotation framework, defined to cover dates, money, percentage and time expressions, names of organizations, locations and persons. Normalized versions of entities are provided to facilitate extraction and comparisons (e.g. for dates and time: values conforming to `xsd:date` and `xsd:time` types, for money: value with ISO currency code).

3. Towards Advanced Processing of Polish

The project targets at practical approach to numerous advanced linguistic issues such as coreference resolution, summarization, machine translation or categorization. Some of them have been recently tackled in other initiatives which can create useful synergies with ATLAS.

3.1. Coreference Resolution

The first attempts of coreference resolution for Polish are being currently carried out within the *Computer-based methods for coreference resolution in Polish texts* project financed by the Polish National Science Centre[1] which intends to create a useful baseline for future experiments with this topic, but also to facilitate comparisons with general coreference resolvers planned for the project such as RARE (Postolache, 2004; Cristea et al., 29 31 May 2002).

The resulting implementation is designed to run either on true mention boundaries (Ogrodniczuk and Kopeć, 2011b) or in an end-to-end manner (Ogrodniczuk and Kopeć, 2011a). The current system uses a few rich rules, corresponding to syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring).

3.2. Text Summarization

Despite intensive worldwide research in the field, just a few general summarization systems for Polish have been implemented so far. The first of them was PolSumm2 (Ciura et al., 2004), a modular, text extraction-based system monitoring inter-sentence relations, anaphors and ellipses. It was then complemented with Lakon (Dudczak, 2007; Dudczak et al., 2008b; Dudczak et al., 2008a; Dudczak et al., 2010), a heuristic summarizer used for testing various sentence selection methods and the extraction-based machine learning system of Joanna Świetlicka (Świetlicka, 2010). The most mature of these approaches can be adapted to ATLAS.

4. References

Szymon Acedański and Konrad Gołuchowski. 2009. A Morphosyntactic Rule-Based Brill Tagger for Polish. In *Recent Advances in Intelligent Information Systems*, pages 67–76, Kraków, Poland, June. Academic Publishing House EXIT.

Szymon Acedański. 2010. A Morphosyntactic Brill Tagger for Inflectional Languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer.

Anelia Belogay, Damir Ćavar, Dan Cristea, Diman Karagiozov, Svetla Koeva, Roumen Nikolov, Maciej Ogrodniczuk, Adam Przepiórkowski, Polivios Raxis, and Cristina Vertan. 2011. i-Publisher, i-Librarian and EU-DocLib – linguistic services for the Web. In *Proceedings of the 8th Practical Applications in Language and Computers Conference (PALC 2011)*, University of Łódź, Poland. 13-15 April 2011.

Marcin Ciura, Damian Grund, Slawomir Kulików, and Nina Suszczanska. 2004. A System to Adapt Techniques of Text Summarizing to Polish. In Ali Okatan, editor, *International Conference on Computational Intelligence*, pages 117–120, Istanbul, Turkey. International Computational Intelligence Society.

Dan Cristea, Oana-Diana Postolache, Gabriela-Eugenia Dima, and Cătălina Barbu. 29-31 May 2002. AR-Engine – a framework for unrestricted co-reference resolution. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*, volume VI, pages 2000–2007, Las Palmas, Gran Canaria, Spain. ELRA.

Łukasz Degórski. 2011. Towards the Lemmatisation of Polish Nominal Syntactic Groups Using a Shallow Grammar. In *Proceedings of the International Joint Conference Security and Intelligent Information Systems*, Warsaw. Institute of Computer Science, Polish Academy of Sciences.

Adam Dudczak, Jerzy Stefanowski, and Dawid Weiss. 2008a. Automatic selection of sentences for Polish newspaper articles (Automatyczna selekcja zdań dla tekstów prasowych w języku polskim, in Polish). Technical Report RA-03/08, Institute of Computing Science, Poznań University of Technology, Poland.

Adam Dudczak, Jerzy Stefanowski, and Dawid Weiss. 2008b. Comparing Performance of Text Summarization Methods on Polish News Articles. In *Proceedings of the International IIS: Intelligent Information Processing and Web Mining Conference*, pages 249–258, Zakopane, Poland.

Adam Dudczak, Jerzy Stefanowski, and Dawid Weiss. 2010. Evaluation of Sentence-Selection Text Summarization Methods on Polish News Articles. *Foundations of Computing and Decision Sciences*, 1(35):27–41.

Adam Dudczak. 2007. Application of selected data exploration methods to summarization of Polish newspaper articles (Zastosowanie wybranych metod eksploracji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego, in Polish). MSc thesis.

Diman Karagiozov, Svetla Koeva, Maciej Ogrodniczuk, and Cristina Vertan. 2011. ATLAS – A Robust Multilingual Platform for the Web. In *Proceedings of the German Society for Computational Linguistics and Language Technology Conference (GSCL 2011)*, Hamburg, Germany. 28-30 September 2011.

[1] Contract number 6505/B/T02/2011/40.

Maciej Ogrodniczuk and Diman Karagiozov. 2011. ATLAS – The Multilingual Language Processing Platform. In *Proceedings of the 27th Conference of the Spanish Society for Natural Language Processing (SEPLN 2011)*, University of Huelva, Spain. 5-7 September 2011.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011a. End-to-end coreference resolution baseline system for Polish. In *Proceedings of the 5th Language & Technology Conference*, Poznań, November. under review.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011b. Rule-based coreference resolution module for Polish. In *Proceedings of DAARC 2011 – the 8th Discourse Anaphora and Anaphor Resolution Colloquium*, Faro, Portugal (to appear), October.

Oana Postolache. 2004. RARE: Robust Anaphora Resolution Engine. MSc thesis.

Adam Przepiórkowski and Aleksander Buczyński. 2007. Spejd: Shallow Parsing and Disambiguation Engine. In Zygmunt Vetulani, editor, *Proceedings of the 3rd Language & Technology Conference*, pages 340–344, Poznań, Poland.

Adam Przepiórkowski and Marcin Woliński. 2003. A Flexemic Tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

Adam Przepiórkowski, Rafal L. Górski, Barbara Lewandowska-Tomaszczyk, and Marek Łaziński. 2008. Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech. ELRA.

Adam Przepiórkowski. 2008. *Powierzchniowe przetwarzanie języka polskiego*. Academic Publishing House EXIT, Warsaw. [In Polish]. B5, 322 pages.

Zygmunt Saloni, Włodzimierz Gruszczyński, Marcin Woliński, and Robert Wołosz. 2007. Grammatical Dictionary of Polish – Presentation by the Authors. *Studies in Polish Linguistics 4, 2007*, pages 5–25. `http://www.ijp-pan.krakow.pl/sipl/saloni.pdf`, see also `http://www.info.univ-tours.fr/~savary/Polonium/Papers/prezentacja-SGJP-Tours.pdf`.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Jakub Waszczuk, Katarzyna Głowińska, Agata Savary, and Adam Przepiórkowski. 2010. Tools and Methodologies for Annotating Syntax and Named Entities in the National Corpus of Polish. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT 2010): Computational Linguistics – Applications (CLA'10)*, pages 531–539, Wisła, Poland. PTI.

Joanna Świetlicka. 2010. Machine learning methods in automatic text summarization (Metody maszynowego uczenia w automatycznym streszczaniu tekstów, in Polish), September. MSc thesis.

Marcin Woliński. 2006. Morfeusz – a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference*, pages 511–520, Wisła, Poland, June.