Expanding the functionalities of the Language Resources Switchboard by integrating a set of tools for the processing of Polish language

Rafał Jaworski
Faculty of Mathematics and Computer Science
Adam Mickiewicz University in Poznań
rjawor@amu.edu.pl

Maciej Ogrodniczuk
Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences
maciej.ogrodniczuk@ipipan.waw.pl

Abstract

This paper presents the Multiservice platform and its integration with the CLARIN Language Resources Switchboard. Multiservice combines a set of offline natural language processing tools for the Polish language. It features, among others, disambiguating tagging, dependency parsing and coreference resolution. A demonstration version of the platform, available online, is also accessible for the CLARIN Language Resources Switchboard (CLRS) users. At CLRS, the user provides a text file, selects one of the predefined processing chains and is automatically redirected to the Multiservice, which is immediately ready to process the request.

1 Introduction

The CLARIN Language Resource Switchboard, described in (Zinn, 2016), here abbreviated CLRS or the Switchboard, is a platform designed to provide an easy access to various natural language processing tools and resources. Its key features include:

- uploading the input text in an electronic textual format;
- automatic recognizing of the language of the text;
- generating a list of tools applicable for processing of the input text;
- redirecting the user's processing request to a chosen tool.

The CLRS relies on integration with external language processing tools, which are configured to accept requests generated by the Switchboard. The required configuration of an external tool is minimal, as it solely consists in implementing a basic web interface, accepting parameters passed by the GET method. The input text file is hosted at CLRS and only its publicly accessible URL is passed to the external tool.

So far, the Switchboard has integrated a variety of language processing tools for different languages, but was somewhat lacking coverage for some of the Slavic languages like Polish. The integration of CLRS and the Multiservice, which incorporates multiple Polish language processing tools itself, aims at improving this coverage.

2 Multiservice

Multiservice (Ogrodniczuk and Lenart, 2012) is a platform created in CLARIN to make offline language processing tools for Polish available as web services and offer their chaining thanks to a common linguistic representation format and asynchronous execution architecture. As of 2016, the toolset comprises several disambiguating taggers (with paragraph-, sentence- and token-level segmentation and morphological analysis): Pantera (Acedański, 2010), WMBT (Radziszewski and Śniatowski, 2011), Concraft (Waszczuk, 2012), WCRFT (Radziszewski, 2013), ensemble tagger PoliTa (Kobyliński, 2014), sentiment analyser Sentipejd (Buczyński and Wawer, 2008), dependency parser (Wróblewska, 2014), shallow parser Spejd (Przepiórkowski and Buczyński, 2007), named entity recognizer Nerf (Waszczuk et al., 2013), two coreference resolvers Ruler (Ogrodniczuk and Kopeć, 2011) and Bartek (Ogrodniczuk et al.,

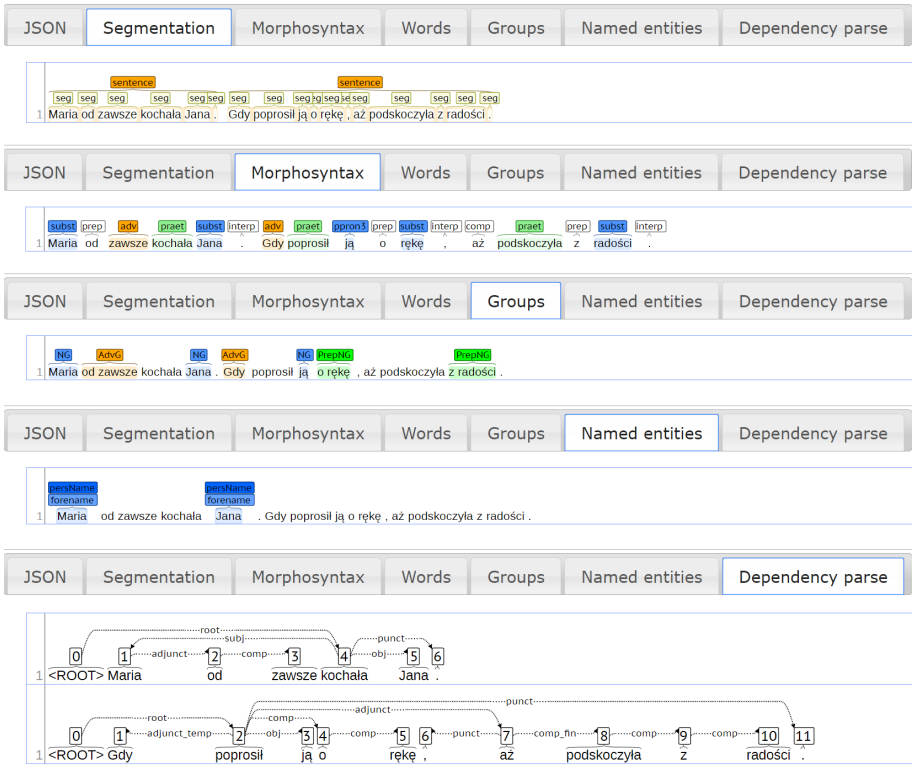

Figure 1: Graphical output of Multiservice: layers of linguistic annotation

2015, chapter 12), OpenTextSummarizer (Rotem, 2003) adjusted for Polish and two other summarization tools: Lakon (Dudczak, 2007), Świetlicka's summarizer (Świetlicka, 2010) and Nicolas (Kopeć, 2016).

Interaction with Multiservice is currently available via a dedicated API available in Java and Python or a web demo of the service (`http://multiservice.nlp.ipipan.waw.pl/`). Both these methods have their drawbacks: programmatical access is in most cases too difficult to use by representatives of the humanities and the web application offers only a brat-based (Stenetorp et al., 2012) interface showing layers of annotations in separate tabs (see Fig. 1) plus a JSON output of results which requires separate retrieval and script-based processing.

An important feature of the Multiservice is the ability to create processing chains consisting of multiple tools. In such chains, the output of one tool becomes the input of another. The chains reflect the logical steps needed to perform specific language processing tasks. For example, a chain to perform shallow parsing would be: Concraft (part-of-speech tagging) → Spejd (shallow parsing).

3 Integration of external tools with the Language Resources Switchboard

The CLARIN Language Resources Switchboard is publicly and freely available at `http://weblicht.sfs.uni-tuebingen.de/clrs/#/` as a web application. Its main goal is to provide an easy-to-use interface for running a variety of natural language processing tools on user-provided files. Once a user uploads a file to the system, its format and language are automatically detected by the CLRS. With the use of that information, a list of applicable tools is generated and presented to the user. When the user selects a tool, a new browser tab is opened and pointed to the URL of the external tool (it is required that all the external tools integrated with the CLRS should expose a web API). The text file uploaded to the CLRS is automatically passed to the external tool, along with optional processing parameters. This is done in order to ensure that the user uploads the data and defines the processing only once. The last step is starting the external tool, which is done using its interface. The entire text processing and results presentation is then performed at the external web application.

This model allows for a relatively simple integration of any language processing tool that is accessible as a web application. From the technical point of view, the text files uploaded to the CLRS are hosted at

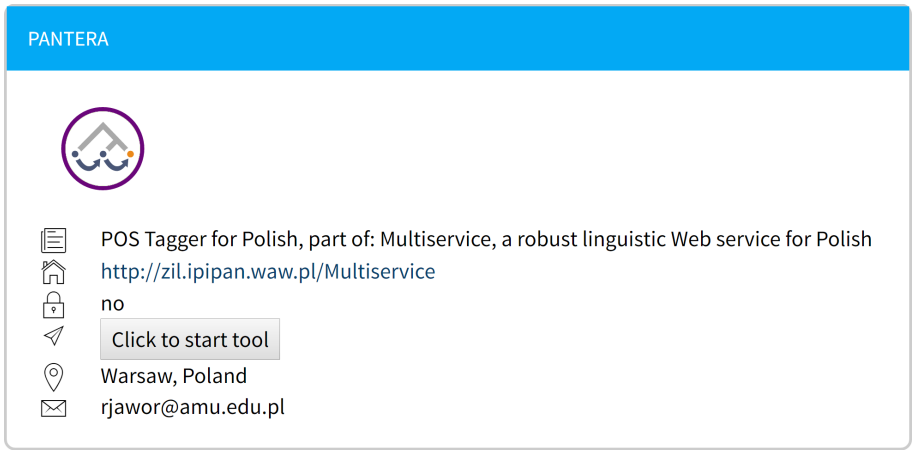

Figure 2: Tool information box for Pantera in the CLRS

the Switchboard's servers. A public URL pointing to the file is then passed to the external tool via a GET parameter. Additional processing options (if any) are also passed this way. Therefore, in the simplest scenario, the only required adjustment to the tool which is to be integrated with the CLRS, is the addition of a mechanism of parsing these GET parameters and setting the processing options in the interface of the tool accordingly.

From the user's perspective, the CLARIN Language Resources Switchboard allows for a prompt and effortless testing of a wide range of processing tools. This may be, for example, a valuable aid in the process of choosing the best tool to use in a production environment.

4 Multiservice tools in CLRS

For the time being, the following Multiservice processing chains have been integrated into the CLRS:

1. Pantera (part-of-speech tagging)
2. Pantera → Spejd (shallow parsing)
3. Pantera → Dependency parser

These chains appear as separate tools in the CLRS. Fig. 2 presents the tool information box for Pantera.

All the above processing tools are configured to operate on plain text files written in Polish. First tests revealed that it might be useful to include in the CLRS an option to detect and, if needed, convert the encoding of the plain text files. Most of the external tools (including the Multiservice) expect files in the UTF-8 encoding, while it is technically possible to upload a file into the CLRS in any encoding. File encoding management in this scenario seems to be an interesting problem to tackle.

5 Conclusions

Integration of Multiservice into the CLARIN Language Resources Switchboard expanded the functionalities of the latter platform by a set of tools designed to process Polish texts. The work on this integration is still ongoing and more tools are expected to be integrated into the Multiservice (and consequently into the CLRS) in the near future.

It is interesting to note that Multiservice and CLRS share the same idea of integrating multiple language processing tools in one platform. The integration of one such platform into another creates a model, where both platforms benefit from each other's development, instead of competing with one another. CLRS expanded its coverage of Slavic languages, while Multiservice significantly improved its visibility in the Web. In biological terms, it may be said that Multiservice and CLARIN Language Resources Switchboard live in a true symbiosis.

References

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.

Aleksander Buczyński and Aleksander Wawer. 2008. Shallow parsing in sentiment analysis of product reviews. In Sandra Kübler, Jakub Piskorski, and Adam Przepiórkowski, editors, *Proceedings of the LREC 2008 Workshop on Partial Parsing: Between Chunking and Deep Parsing*, pages 14–18, Marrakech. ELRA.

Adam Dudczak. 2007. Zastosowanie wybranych metod eksploracji danych do tworzenia streszczeń tekstów prasowych dla języka polskiego. Master's thesis.

Łukasz Kobyliński. 2014. PoliTa: A multitagger for Polish. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2949–2954, Reykjavík, Iceland. ELRA.

Mateusz Kopeć. 2016. Nicolas Summarizer. [on-line] `http://zil.ipipan.waw.pl/Nicolas`.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011. Rule-based coreference resolution module for Polish. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 191–200, Faro, Portugal.

Maciej Ogrodniczuk and Michał Lenart. 2012. Web Service integration platform for Polish linguistic resources. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1164–1168, Istanbul, Turkey. ELRA.

Maciej Ogrodniczuk, Katarzyna Głowińska, Mateusz Kopeć, Agata Savary, and Magdalena Zawisławska. 2015. *Coreference in Polish: Annotation, Resolution and Evaluation*. Walter De Gruyter.

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

Adam Radziszewski and Tomasz Śniatowski. 2011. A memory-based tagger for Polish. In Zygmunt Vetulani, editor, *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2011)*, pages 29–36.

Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, and M. Niezgódka, editors, *Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions*. Springer Verlag.

Nadav Rotem. 2003. The Open Text Summarizer. [on-line] `http://libots.sourceforge.net/`.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joanna Świetlicka. 2010. Metody maszynowego uczenia w automatycznym streszczaniu tekstów. Master's thesis.

Jakub Waszczuk, Katarzyna Głowińska, Agata Savary, Adam Przepiórkowski, and Michał Lenart. 2013. Annotation tools for syntax and named entities in the National Corpus of Polish. *International Journal of Data Mining, Modelling and Management*, 5(2):103–122.

Jakub Waszczuk. 2012. Harnessing the CRF complexity with domain-specific constraints. The case of morphosyntactic tagging of a highly inflected language. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2789–2804, Mumbai, India.

Alina Wróblewska. 2014. *Polish Dependency Parser Trained on an Automatically Induced Dependency Bank*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Claus Zinn. 2016. The CLARIN Language Resource Switchboard. [on-line] `https://www.clarin.eu/sites/default/files/zinn-CLARIN2016_paper_26.pdf`.