Polish Language Processing Chains for Multilingual Information Systems*

Maciej Ogrodniczuk and Adam Przepiórkowski

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, Warsaw, Poland

maciej.ogrodniczuk@ipipan.waw.pl
adam.przepiorkowski@ipipan.waw.pl

Abstract. The ATLAS project, started in March 2010, intends to create a multilingual language processing framework integrating the common set of linguistic tools for a group of European languages, among them Polish. The chained tools producing multi-level UIMA-encoded annotation of texts can be used by NLP applications for complex language-intensive operations such as automated categorization, information extraction, machine translation or summarization.
This paper concentrates on applications of ATLAS language processing chains to multilingual information systems, with particular interest in processing Polish. Inflectional characteristics of this language offers the possibility to comment on a few more advanced functions such as multiword unit lemmatisation, vital for real-life presentation of extracted phrases. Several sample applications using the NLP chain are also presented.

1 Introduction

The ATLAS project[1] [8] offered the possibility to test interoperability of several NLP tools for Polish together with other project languages — Bulgarian, English, German, Greek and Romanian — in multilingual information systems. After selecting the common integration and annotation framework, the language tools capable of providing necessary linguistic information have been evaluated and adapted to form ready-to-use processing chains. Currently three linguistically-aware online services make use of this functionality: i-Publisher (Web-based content management system), i-Librarian (a digital library of scientific works) and EUDocLib/PLDocLib sites (for browsing and searching through EUR-LEX documents and, respectively, acts of Polish Parliament).

* 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).

[1] See http://www.atlasproject.eu/ for a detailed information about the project.

2 Language Processing Chains

Combining diverse tools into a single framework required selection of an integration platform for linguistic annotation. Unstructured Information Management Architecture (UIMA) had been chosen from among other reputable architectures (such as General Architecture for Text Engineering), mainly because its potential for scalability achievable by decomposition of language processing applications into components replicable over a cluster of network nodes.

The NLP tools have been integrated into UIMA annotation framework by means of wrapping them into UIMA-compatible primitive engines, chained into aggregate engines. In certain cases it required their technical adaptation to continuous use (without frequent loading of models, processing rules, etc.) and enforcing their thread-safety.

Besides, wrappers were designed to maintain a uniform type system developed for ATLAS, with document-level and text-level properties, the latter comprising annotations of paragraphs, tokens (POS tags and morphosyntactic categories), noun phrases (with semantic heads) and named entities.

3 Polish Language Tools

A number of existing tools for the processing of Polish were used within 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.

3.1 Segmentation and Morphosyntactic Tagging

Sentence- and token-level segmentation information is provided by Morfeusz [16], also a lemmatizer and morphological analyzer for Polish. It uses positional tags starting with POS information followed by values of morphosyntactic categories corresponding to the given part of speech [12]. Current version of the tool, Morfeusz SGJP, is based on linguistic data coming from The Grammatical Dictionary of Polish [13].

Morphosyntactic disambiguation is performed by Pantera [1], a rule-based tagger of Polish using 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 the free word order nature of Polish, the original set of rule templates as proposed by Brill has been extended to cover larger contexts.

3.2 Noun Phrase Extraction and Multi-Word Expression Lemmatisation

Noun phrases are identifier by Spejd [2], a shallow parsing engine using cascade grammars, able to co-operate with various taggers of Polish, including Pantera (see above). In Spejd parsing rules are defined using a cascade of regular grammars which match against orthographic forms, base 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), etc. Users may provide custom rules or may use one of the provided sample rule sets.

Apart from identifying noun phrases, the Spejd grammar of Polish [6] created within the National Corpus of Polish [11] is the basis for the nominal groups lemmatiser [4] developed throughout the project which combines the lemmatisation task with shallow parsing. The parsing structures are used in lemmatisation schemata written separately for each grammar rule and operating on the matched strings and structure.

3.3 Named Entity Recognition

Identification of named entities is supported by NERF tool [14] – a statistical CRF-based named entity recognizer trained over 1-million manually annotated subcorpus of the National Corpus of Polish and successfully used in the process of automated annotation of its total 1,5 billion segments.

NERF annotation model is consistent with general requirements of the ATLAS 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. values conforming to `xsd:date` and `xsd:time` types for date/time expressions and ISO currency codes for money expressions).

4 Current Work

The project targets at practical approach to numerous advanced linguistic issues such as coreference resolution, summarization, machine translation or categorization, often creating synergies with other ongoing initiatives. Initial versions of tools providing the above-mentioned operations are scheduled to be integrated by the time of NLDB 2012.

4.1 Text Summarization and Coreference Resolution

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 [3], a modular, text extraction-based system monitoring inter-sentence

relations, anaphors and ellipses. In 2007 Lakon [5], a heuristic summarizer used for testing various sentence selection methods was implemented and in 2010 – the extraction-based machine learning system of Świetlicka [15]. They are currently being evaluated against shallow language-independent clause- and marker-based extractive summarizer implemented by Romanian partners and adapted to ATLAS.

To achieve better results, a coreference resolution module is currently being integrated into the summarization engine with two approaches being evaluated. The first of them is a language-neutral coreference resolver RARE [10], the second one is the results of the first attempts of coreference resolution for Polish carried out within the *Computer-based methods for coreference resolution in Polish texts* project financed by the Polish National Science Centre. Currently used end-to-end implementation [9] adopts a rich rule-based approach, integrating syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring).

4.2 Categorization and Machine Translation

General language-independent categorization tools for heterogeneous domains have been integrated into ATLAS. The engine employs different categorization algorithms, such as Naïve Bayesian, relative entropy, Class-Feature Centroid, Support Vector Machines and Latent Dirichlet Allocation, with results consolidated by a voting system.

The machine translation engine currently being evaluated combines an example-based component with a statistical, domain-factored approach powered by Moses [7]. After the most appropriate translation model and example-based sub-component are selected based on the results from categorisation engine, the translation database and the statistical component are used to provide the translation output. Before the engine reaches its maturity, a third-party solution using Microsoft Bing is being used by the demo interfaces described below.

5 Polish Linguistic Chain-based NLP Demo Applications

The UIMA linguistic annotation chains for Polish have been tested together with higher-level linguistic functions such as summarization, categorization and machine translation in a several sample Web site interfaces implemented as demonstrations of the technology.

5.1 i-Librarian and EUDocLib

i-Librarian (`http://www.i-librarian.eu/`) is a free online library that assists authors, students, young researchers, scholars, librarians and executives to easily create, organise and publish various types of documents; EUDocLib (`http://eudoclib.atlasproject.eu/`) is a publicly accessible repository of EU

documents from the EUR-LEX collection which provides easier access to relevant documents in the user's language.

Both sites are capable of processing documents in supported languages in order to automatically categorize, summarize and annotate content with important noun phrases and named entities. They also provide annotation-based content navigation (such as list of similar documents) and machine-translated excerpts of documents used for document categorization and clustering.

5.2 PLDocLib

PLDocLib (`http://www.atlasproject.eu/pl/`) is a language processing chain-powered Web site offering full-text search and category-based browsing of around 1000 acts of Polish Sejm. For each document a set of recognized named entities, automatically clustered important noun phrases (with their weights) and a list of similar documents is produced. For presentation, base forms of multiword units are generated and manually assigned categories (retrieved from the document source) are used.

6 Conclusions

While the number of resources and tools for European languages grows rapidly, their interoperability leaves much to be wished for, which hinders development of multilingual information systems. In this paper we reported on a practical exercise in making Polish language processing tools interoperable.

At the technical level, the interoperability is ensured by integrating the tools within the UIMA platform. In order to do that, the tools themselves did not need to be substantially modified, but appropriate wrappers around them had to be implemented.

At the more interesting linguistic level, the tools were created with the intention of using them in sync, although they had never before been combined into a processing chain like the one described here. In particular, the Pantera tagger assumes positional tagsets of the kind employed by the morphological analyser Morfeusz, the shallow parsing system Spejd can in principle deal with any morphosyntactic tagsets, but frontends exist reading the format produced by tools like Pantera and, moreover, the grammar developed within the National Corpus of Polish assumes the same tagset (by now standard in Polish NLP). Although Polish NER tools were developed in the same project, they have not been used together with the shallow parser so far.

The deployment of these tools in ATLAS applications serves as an important proof of concept that the intended interoperability of these tools is indeed possible and relatively straightforward.

References

1. Acedański, S.: A Morphosyntactic Brill Tagger for Inflectional Languages. In: Loftsson, H., Rögnvaldsson, E., Helgadóttir, S. (eds.) Advances in Natural Language

Processing. Lecture Notes in Computer Science, vol. 6233, pp. 3–14. Springer (2010)
2. Buczyński, A., Przepiórkowski, A.: Spejd: A shallow processing and morphological disambiguation tool. In: Vetulani, Z., Uszkoreit, H. (eds.) Human Language Technology: Challenges of the Information Society, Lecture Notes in Artificial Intelligence, vol. 5603, pp. 131–141. Berlin (2009)
3. Ciura, M., Grund, D., Kulików, S., Suszczanska, N.: A System to Adapt Techniques of Text Summarizing to Polish. In: Okatan, A. (ed.) Proceedings of the International Conference on Computational Intelligence (ICCI 2004). pp. 117–120. International Computational Intelligence Society, Istanbul, Turkey (2004)
4. Degórski, L.: Towards the Lemmatisation of Polish Nominal Syntactic Groups Using a Shallow Grammar. In: Bouvry, P., Kłopotek, M.A., Leprevost, F., Marciniak, M., Mykowiecka, A., Rybiński, H. (eds.) Security and Intelligent Information Systems: International Joint Confererence, SIIS 2011. Revised Selected Papers. Lecture Notes in Computer Science, vol. 7053. Springer-Verlag (2011)
5. Dudczak, A., Stefanowski, J., Weiss, D.: Evaluation of Sentence-Selection Text Summarization Methods on Polish News Articles. Foundations of Computing and Decision Sciences 1(35), 27–41 (2010)
6. Głowińska, K., Przepiórkowski, A.: The design of syntactic annotation levels in the National Corpus of Polish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010. ELRA, Valletta, Malta (2010)
7. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 177–180. ACL '07 (2007)
8. Ogrodniczuk, M., Karagiozov, D.: ATLAS — The Multilingual Language Processing Platform. Procesamiento del Lenguaje Natural 47, 241–248 (2011)
9. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for Polish. In: Proceedings of the 5th Language & Technology Conference (LTC 2011). pp. 167–171. Poznań, Poland (2011)
10. Postolache, O.: RARE: Robust Anaphora Resolution Engine. Master's thesis, University of Iasi (2004)
11. Przepiórkowski, A., Górski, R.L., Łaziński, M., Pęzik, P.: Recent developments in the National Corpus of Polish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010. ELRA, Valletta, Malta (2010)
12. Przepiórkowski, A., Woliński, M.: A Flexemic Tagset for Polish. In: Proceedings of Morphological Processing of Slavic Languages, EACL 2003 (2003)
13. Saloni, Z., Gruszczyński, W., Woliński, M., Wołosz, R.: Grammatical Dictionary of Polish – Presentation by the Authors. Studies in Polish Linguistics 4, 2007 pp. 5–25 (2007)
14. Savary, A., Waszczuk, J., Przepiórkowski, A.: Towards the Annotation of Named Entities in the National Corpus of Polish. In: Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010. Valletta, Malta (2010), ELRA
15. Świetlicka, J.: Machine learning methods in automatic text summarization (in Polish). Master's thesis, Warsaw University, Poland (2010)
16. Woliński, M.: Morfeusz – a practical tool for the morphological analysis of Polish. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference. pp. 511–520. Wisła, Poland (2006)