Rule-based coreference resolution module for Polish*

Maciej Ogrodniczuk and Mateusz Kopeć

Institute of Computer Science, Polish Academy of Sciences

Abstract. This paper presents the results of the first attempt of coreference resolution for Polish running on true mention boundaries and using 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). The results are compared to human annotation and presented in four sets: with two common baselines: all-singletons/head-match, and two slightly more complex settings with four and five rules.

Keywords: coreference resolution of Polish, Polish anaphora resolution

1 Introduction

Although few anaphora resolution attempts were already made for Polish (see e.g. [7], [8], [9]) they were either purely theoretical or pronoun-limited. This paper presents the first coreference resolution module for Polish, intended to provide starting ground for further experiments and generate the reference baseline to be compared with future more advanced rule-based and statistical coreference resolvers. The scope of the resolution is limited to identity-of-reference direct nominal coreference.

The module design follows Haghighi and Klein's approach [4] by building on the richness of important characteristics rather than multitude of weak features. For languages such as Polish which still lacks advanced discourse processing tools, this approach seems very promising also because of practical reasons.

Additional intention of this attempt is gathering experience for the next phases of recently started project *Computer-based methods for coreference resolution in Polish texts* which tasks also include creation of the corpus of Polish manually annotated with various types of coreference.

2 System Description

The implemented module uses standard best-first entity-based model based on syntactic constraints (elimination of nested nominal groups), syntactic filters

* The work reported here was carried out within the *Computer-based methods for coreference resolution in Polish texts* project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40).

(elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring). Syntactic properties are obtained from Spejd and its morphological component Morfeusz SGJP which produce NP chunks with detailed morphosyntatic information. Semantic properties are currently based on Polish WordNet (all tools see Sec. 2.1).

2.1 External Resources and Tools

Several basic tools for providing disambiguated morphosyntactic information, syntactic groups and named entities were used by the module.

Morfeusz Morfeusz [22] is a morphological analyzer for Polish. It uses a positional tags starting with POS information followed by values of morphosyntactic categories corresponding to the given part of speech [15]. Current version of the tool, Morfeusz SGJP, is based on linguistic data coming from The Grammatical Dictionary of Polish [18].

The tool is available at `http://sgjp.pl/morfeusz/index.html` and is distributed under the terms of the GNU AGPL 3[1].

Pantera Pantera [1,2] 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. The achieved error rate amounts to 10.8%, but the tagger is currently under active development.

Spejd Spejd [12,13] is an engine for shallow parsing using cascade grammars, able to co-operate with TaKIPI for tokenization, segmentation, lemmatization 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 also available as a separate online service at `http://chopin.ipipan.waw.pl:8081/spejdws/services/SpejdService?wsdl`.

[1] Demo online version of the older variant of the morphological analyser, Morfeusz SiAT, is still available at `http://sgjp.pl/demo/morfeusz`.

NER Tools NER [19,20] is a statistical CRF-based named entity recognition tool trained over 1-million manually annotated subcorpus of the National Corpus of Polish [14] and successfully used in the process of automated annotation of its total 1 billion segments.

The annotation scope is defined to cover personal names, geographical names, names of organizations and institutions, words related to the above categories (relational adjectives, names of inhabitants and organization members) and basic temporal expressions. The taxonomy of annotation features medium degree of granularity with 2 levels for e.g. place names or personal names (with forenames, surnames etc.)

Polish Wordnet Polish WordNet [11][2] is a network of lexical-semantic relations, an electronic thesaurus with a structure modelled on that of the Princeton WordNet and those constructed in the EuroWordNet project. Polish WordNet describes the meaning of a lexical unit of one or more words by placing this unit in a network of links which represent such relations as synonymy, hypernyny, meronymy etc.

To reduce the cost of the project, Polish WordNet has been built semi-automatically. Lexical relations were automatically recognized in large corpora of Polish and suggested to linguists/lexicographers via a graphical interface.

2.2 Scoring

For a new mention candidate its compatibility with all previously constructed chains is calculated and the best cluster is selected (only when the score exceeds the threshold value, currently 0.5). When more than one chain results in the best score, the one containing the closest mention is selected. The compatibility of the candidate mention and a given chain is defined as the maximum of the compatibility scores of the mention tested against each of the chain's mentions.

The scoring of compatibility of two mentions starts with 0.5 value for the mention being investigated (which corresponds to equal chances of compatibility/incompatibility with the chain) and consists in applying the 5 rich rules:

1. *gender/number rule* eliminates syntactically incompatible matches, i.e. prevent mentions with different gender or number to be marked as coreferent,
2. *including rule* eliminates nested groups, not allowing to put two mentions having a non-empty intersection in one cluster,
3. *lemma rule*, turned on for nominal groups only (not pronouns), promotes matches with identical heads and lowers the total score for incompatible heads,
4. *wordnet rule*, valid only for nominal groups which have their wordnet representation, increases the score when the topic set containing synonyms, hyperonyms, alternyms and fuzzynyms intersect with more than 3 entries, and decreases it otherwise,

[2] Also known as plWordNet or Słowosieć; see `http://plwordnet.pwr.wroc.pl/wordnet/`.

```
Stare        stary         adj     -    -   -   -        (22
Amerykanki   Amerykanka    subst   nom  pl  f   head:22  22)NG
biegają      biegać        fin     -    -   -   -
po           po            prep    -    -   -   -
dziedzińcu   dziedziniec   subst   loc  sg  m3  -        (24)NG
,            ,             interp  -    -   -   -
po           po            prep    -    -   -   -
sypialni     sypialnia     subst   loc  sg  f   head:26  (26
emira        emir          subst   gen  sg  m1  -        (10)NG|26)NG
,            ,             interp  -    -   -   -
fotografują  fotografować  fin     -    -   -   -
,            ,             interp  -    -   -   -
zaglądają    zaglądać      fin     -    -   -   -
w            w             prep    -    -   -   -
głąb         głąb          subst   acc  sg  m3  -
lochów       loch          subst   gen  pl  m3  -        (30)NG
.            .             interp  -    -   -   -
```

Fig. 1. Input format example

5. *pronoun rule*, valid for pronouns only, increases the score of matching pronoun with any other mention, because pronouns mostly appear in text after a non-pronoun coreferent and therefore should be a part of a chain (it also lowers the score for incompatible first and second-person pronouns, because they do sometimes occur in texts without non-pronoun coreferents).

2.3 Input and output format

Implemented module requires the text to be annotated to be provided in a specific format, which is based on the format used during the coreference resolution competition during the Semeval conference [16]. It's example can be seen in figure 1.

Rows in input file correspond to single tokens in text and each row consists of a number of attributes for a particular token, separated by whitespace. First value is orthographic form of the token, followed by it's lemmatized form, part of speech, case, number and gender. Next values are optional – they describe the mention layer of the document. If the token is a part of a mention and it's head of it, there is a value indicating it, for example `head:0`, if the id of this mention is `0`. Last value describes what mentions this token belongs to, using bracket notation almost the same as it was during Semeval. The only modification is the obligation to mark the type of mention after its closing bracket. It can be either `NG` (noun group), `NE` (named entity) or `P` (pronoun).

For example, phrase `sypialni emira` (emir's bedroom) is a mention with id 26 of type `NG` (noun group) and it's head is `sypialni` (bedroom). There is also a nested mention `emira`, with id 10 and type of noun group (notice that there is no need to mark the head in a one-token mention).

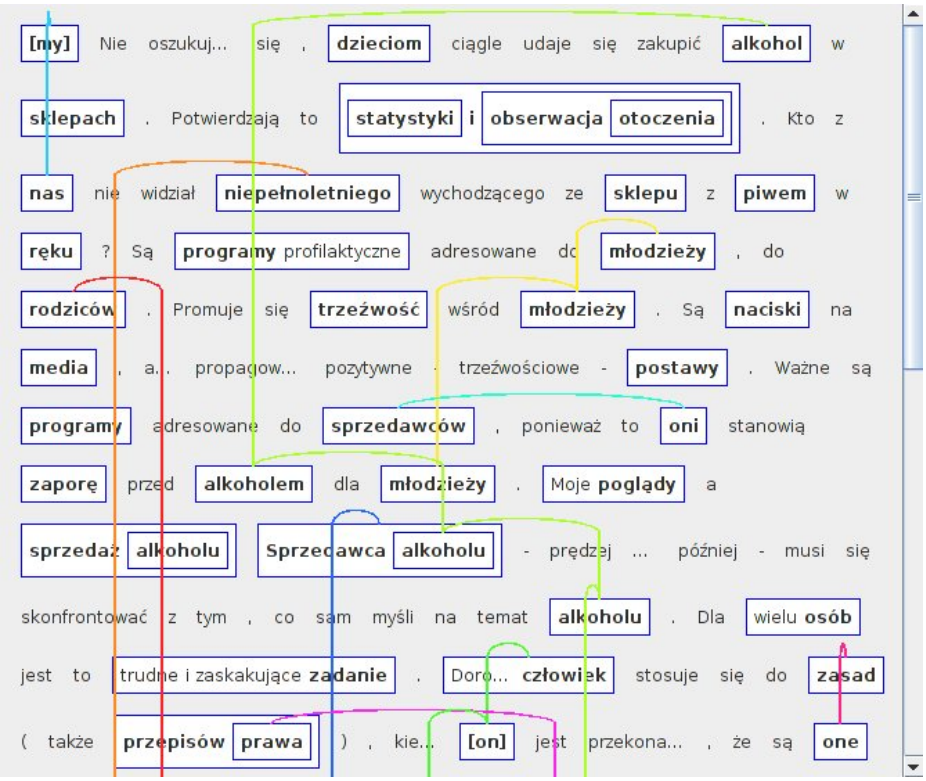

Fig. 2. Example visualization of coreference relations

Output format is the same as the input format, except the types of mentions are removed and the ids of mentions in one mention chain are unified to one value (which means that head information is no longer valid). This format allows output files to be processed directly using evaluation script provided by Semeval task organisers.

2.4 Visualization

As part of the project, a very simple coreference visualization engine was developed. Example of its out output is presented in figure 2. Mentions are repesented as white boxes with blue border, coreference chains are marked with color lines (different colors do not have any special meaning – they are used to make visualization more readable).

3 Data Sets and Evaluation

Evaluation data came from the balanced part of the National Corpus of Polish which provided 50 randomly selected texts. A small sample of 20 subsequent

Mention chain length	1	2	3	4	5	6	7	8	...
Number of chains	1079	88	43	20	9	6	3	2	...
Mention chain length	...	9	10	11	12	15	22	27	Any
Number of chains	...	2	5	1	1	1	1	1	1262

Table 1. Mention chains size statistics

sentences was extracted from each of the texts. 35 samples were used as development data and 15 samples as testing data.

For evaluation, all the data files have been automatically pre-processed with noun phrase chunker (Spejd) and presented to the linguist who verified and corrected mention borders and their morphosyntactic descriptions. The results of this verification were then used as input data for both the manual chain annotation (resulting in producing the gold standard) and the automated coreference module.

The module was designed to provide environment for testing coreference rule sets, which facilitated creating two common variations: the all-singletons and head-match baselines plus slightly more complex, although still very straighforward settings, with all 5 rules described above and – additionally – another run with smaller set of four rules (wordnet rule turned off) to illustrate an interesting discovery.

Although it must be noted that using gold mention boundaries creates somewhat unrealistic running conditions as compared to end-to-end systems, it allows for clear separation of mention detection and coreference resolution which adds to the clarity of the proposed solution.

3.1 Data statistics

All samples contained 6498 mentions, ≈1000 sentences and ≈20000 tokens, the average mention had the length of 1.9 tokens. The number of sentences and tokens were not equal to 1000 and 20000 respectively because of the fact, that the 20-sentence samples were selected based on automatical tokenization and sentence-splitting. In few cases, the sentence boundaries were not selected properly and were corrected by a linguist.

The testing data consisting of 15 text samples contained 1737 mentions, which formed 1262 mention chains. Most of the mention chains consisted of only one mention, which is a standard ratio for a 20-sentence discourse (since most of the entities are referenced only once). The average size of mention chain was 1.37 mentions; detailed statistics are presented in the table 1.

3.2 Evaluation metrics

Four implemented rule sets were evaluated against four well-known coreference resolution evaluation metrics: MUC [6], CEAF [5], B^3 [3] and BLANC [17].

System type	MUC			CEAF		
	R	**P**	**F1**	**R**	**P**	**F1**
All-sing.	–			93.10%	67.64%	78.35%
All-sing. + head m.	50.73%	61.16%	55.46%	84.22%	79.14%	81.60%
5 rules	75.36%	59.46%	66.48%	78.62%	87.42%	82.79%
4 rules (no wordnet)	74.73%	65.13%	69.60%	83.45%	88.36%	85.84%
	B^3			**BLANC**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-sing.	72.65%	100.00%	84.16%	50.00%	49.18%	49.58%
All-sing. + head m.	84.17%	90.05%	87.01%	69.64%	84.54%	74.97%
5 rules	90.56%	82.56%	86.37%	81.99%	78.39%	80.08%
4 rules (no wordnet)	90.35%	86.66%	88.47%	81.94%	83.92%	82.90%

Table 2. Experimental results

Because the output of the system was generated in Semeval format, there was no need to implement new comparator as we were able to use script provided by organisers of the Task 1: Coreference Resolution in Multiple Language from SemEval-2010 competition (see [16]). This script is able to compare two files (both encoded in Semeval format) - one containing golden standard annotations and the other being created by the system under test. The output has the results for each of the four metrics mentioned earlier, both in terms of precision and recall, as well as F1 measure.

4 Experimental Results

The formal experimental results are presented in table 2.

The most interesting finding is that the wordnet rule, although usually adding to the recall, lowers the precision of the score, which can result from the fact that the topic sets can contain very occasionally used meanings producing false positives in unexpected contexts. For instance, word *strona* (part) can be marked as coreferent to *ziemia* (land), because there is an expression in Polish *„strony ojczyste”* („native land”) and it appears in the wordnet. There is a need to develop a more sophisticated method of using the wordnet, which should have to do with word sense disambiguation.

5 Towards End-to-end Coreference Resolution System for Polish

The most recent yet still preliminary results concerning extension of the module into unsupervised coreference resolver for Polish were described in [10]. The annotation produced by the chain-detecion module presented in this article was preceded with a separate module responsible for processing raw text and (after

enriching it using existing natural language processing tools for Polish language) automatic detection of the mentions.

Raw text is part-of-speech-tagged with Pantera, shallow parsed with Spejd and processed by the Named Entity Recognizer. Obtained information from each of this tools is then used to collect mention boundaries. The candidates for mentions are all the nouns and pronouns from the morphosyntactical level, all the nominal groups from the shallow parsing results, and finally all the named entities. Conflicts between the candidates, as well as redundancies are resolved heuristically.

As the final result of mention detection, text is saved in SemEval-like format (described earlier as the input format for coreference resolution module), persisting some of the morphosyntactical information and mention boundaries, as well as mention head indication, where applicable.

The main room for improvement for automatic mention detection module is an addition of zero anaphora detector, as it was proven in [10] that zero anaphoras play a significant part of the final result measured with all evaluation metrics.

6 Conclusions and Further Work

The presented approach offers a useful yet easy to implement baseline for further work and is currently, despite its limited scope, the only available coreference resolution module for Polish.

Further planned tasks include broadening the range of represented coreference types, refinement of the Spejd grammar used for mention identification, machine learning experiments and expanding the feature base with other rich syntactic and semantic features (e.g. by using the results of deep parsing of Polish with Świgra [21] as well as information extracted from Polish Wikipedia and other available fact bases). The results of this process are also intended to create synergy with ATLAS project[3] where anaphora resolution module is planned to be integrated in the summarization component.

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), http://ripper.dasie.mimuw.edu.pl/~accek/homepage/wp-content/papercite-data/pdf/ace10.pdf
2. Acedański, S., Gołuchowski, K.: A Morphosyntactic Rule-Based Brill Tagger for Polish. In: Recent Advances in Intelligent Information Systems. pp. 67–76. Academic Publishing House EXIT, Kraków, Poland (June

[3] Applied Technology for Language-Aided CMS co-funded by the European Commission under the Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467).

2009), http://ripper.dasie.mimuw.edu.pl/~accek/homepage/wp-content/papercite-data/pdf/acegol09.pdf
3. Bagga, A., Baldwin, B.: Algorithms for scoring coreference chains. In: The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference. pp. 563–566 (1998)
4. Haghighi, A., Klein, D.: Simple coreference resolution with rich syntactic and semantic features. In: EMNLP'09. pp. 1152–1161 (2009)
5. Luo, X.: On Coreference Resolution Performance Metrics pp. 25–32 (2005)
6. Marc Vilain and John Burger and John Aberdeen and Dennis Connolly and Lynette Hirschman: A Model-Theoretic Coreference Scoring Scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6). pp. 45–52 (November 1995)
7. Marciniak, M.: Anaphor Binding in Polish. Theory and Implementation. In: Proceedings of DAARC 2002 – the 4th Discourse Anaphora and Anaphor Resolution Colloquium. Lisbon (2002)
8. Matysiak, I.: Information Extraction Systems and Nominal Anaphora Analysis Needs. In: Proceedings of the International Multiconference on Computer Science and Information Technology. pp. 183—-192 (2007)
9. Mitkov, R., Belguith, L., Styś, M.: Multilingual Robust Anaphora Resolution. In: Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP-3). pp. 7–16. Granada, Spain (1998)
10. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for polish. In: Proceedings of the 5th Language & Technology Conference. Poznań (2011), under review
11. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej (2009), http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf
12. Przepiórkowski, A.: Powierzchniowe przetwarzanie języka polskiego. Academic Publishing House EXIT, Warsaw (2008), [In Polish]. B5, 322 pages
13. Przepiórkowski, A., Buczyński, A.: Spejd: Shallow parsing and disambiguation engine. In: Proceedings of the 3rd Language & Technology Conference. Poznań (2007), http://nlp.ipipan.waw.pl/~adamp/Papers/2007-ltc-spade/Spade.pdf
14. Przepiórkowski, A., Górski, R.L., Lewandowska-Tomaszczyk, B., Łaziński, M.: Towards the National Corpus of Polish. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008. ELRA, Marrakech (2008)
15. Przepiórkowski, A., Woliński, M.: A Flexemic Tagset for Polish. In: Proceedings of *Morphological Processing of Slavic Languages*, EACL 2003 (2003), http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf
16. Recasens, M., Marquez, L., Taulé, M., Martí, M.A., Hoste, V., Poesio, M., Versley, Y.: SemEval-2010 Task 1: Coreference Resolution in Multiple Languages, pp. 70–75. No. July, Association for Computational Linguistics (2010), http://www.aclweb.org/anthology/S10-1001
17. Recasens, Marta and Eduard Hovy: BLANC: Implementing the Rand index for coreference evaluation. Natural Language Engineering pp. 1–26 (2010)
18. 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), http://www.ijp-pan.krakow.pl/sipl/saloni.pdf, see also http://www.info.univ-tours.fr/~savary/Polonium/Papers/prezentacja-SGJP-Tours.pdf

19. Savary, A., Waszczuk, J., Przepiórkowski, A.: 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 (2010), eLRA
20. Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A.: 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). pp. 531–539. Wisła, Poland (2010), pTI
21. Woliński, M.: Computer-aided verification of Świdziński's grammar. PhD. diss., Warsaw (2004), `http://www.ipipan.waw.pl/~wolinski/publ/mw-phd.pdf`, [In Polish]. PhD dissertation. Institute of Computer Science, Polish Academy of Sciences
22. 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 (June 2006)