End-to-end coreference resolution baseline system for Polish

Maciej Ogrodniczuk, Mateusz Kopeć

Institute of Computer Science
Polish Academy of Sciences
ul. Ordona 21, Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl, mkopec87@gmail.com

Abstract

The following paper presents the results of the first attempt of coreference resolution for Polish, intended to create a useful baseline for future experiments with this topic. The resulting implementation is designed to run either on true mention boundaries (discovering coreference chains between them) or in an end-to-end manner, performing their automatic detection as the first step. The 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). Results are evaluated against human annotation for two commonly used baseline variants of the resolver (all-singletons/head-match) and two target rule-based settings. The best working method is analysed, showing simple statistics about the two classes of errors made by the system.

Keywords: coreference resolution, anaphora resolution, mention identification

1. Introduction

A few early computational anaphora resolution approaches were made for Polish in late 1990s and 2000s (see e.g. (Mitkov et al., 1998), (Marciniak, 2002), (Matysiak, 2007)), but their scope was rather limited. In our best judgement, far more extensive topic of coreference resolution of Polish was never covered by an end-to-end solution. This paper presents the first coreference resolution system for Polish, intended to provide a point of departure for further experiments and generate the reference baseline to be compared with future more advanced rule-based and statistical coreference resolvers.

1.1. Noun phrase coreference resolution task

The task of noun phrase (NP) coreference resolution is usually defined as "determining which NPs in a text or dialogue refer to the same real-world entity" (Ng, 2010) and is usually implemented as a two-step process:

1. identification of mentions (in current task: a group of adjacent words having nominal head, e.g. pronouns, proper nouns, nominal groups etc.),
2. building coreference chains with mentions having identical referent.

A given fragment of text can produce several nested mentions in the form of NPs with different semantic heads which correspond to different real-world objects. When a mention is identified, the most descriptive sequence of words is stored (provided that it does not contain verbs in finite forms). For example, the fragment *dyrektor departamentu firmy* (company department director) contains 3 mentions:

1. the whole phrase *dyrektor departamentu firmy*,
2. the subphrase referring to the company department (departamentu firmy), and
3. the subphrase referring to the company (firmy).

Sequence of words such as *dyrektor departamentu* (department director) is not marked because there is a longer (and more informative) sequence (the whole phrase) sharing the same semantic head.

Such definition, even though it restricts the wide set of referring structures to nominal constructs, is most often further constrained to cover *identity-of-reference* relations only (excluding identity-of-sense anaphora, bound or bridging anaphora etc.). Having stated that, we currently limit the scope of the resolution in the presented system to identity-of-reference direct nominal coreference.

1.2. Methodology

The module design follows Haghighi and Klein's approach (Haghighi and Klein, 2009) by using a few rich linguistic features rather than tens of less important characteristics. For languages such as Polish which still lacks advanced dis course processing tools, this approach seems very promising also because of practical reasons.

The described solution is an extension to the coreference resolution module running on gold mentions only, presented in (Ogrodniczuk and Kopeć, 2011).

2. System description

The implemented system can be divided into two modules. First module is responsible for processing raw text and (after enriching it using existing natural language processing tools for Polish language) automatic detection of the mentions.

Second module has the ability to find chains (or clusters) of mentions, such that each mention in a chain is referring to the same real-life entity. Of course the accuracy of this assignment is not perfect – therefore this module provides also an interface to calculate various coreference resolution evaluation measures. The input to this module consists of two corpora: one, which already has detected mentions, but not the mention chains, and the other, which is the gold standard for the evaluation procedure, and should have also correctly tagged clusters.

Presented modularity allows to evaluate two scenarios. First, using them both in single pipeline, it allows to evaluate a system automatically detecting mentions and mention chains from raw text. Second, when utilising only the group-finding part, and providing it hand-tagged mention boundaries (but not mention groups!) it allows the measurement of the performance of coreference resolution itself, in separation from the mention-detection part of the process.

2.1. Mention detection

The processing of raw text begins with part-of-speech tagging with Pantera (Acedański and Gołuchowski, 2009; Acedański, 2010). Then text is shallow parsed with Spejd (Przepórkowski and Buczyński, 2007) and its morphological component Morfeusz SGJP (Woliński, 2006). Last step is finding of Named Entities performed by NER (Savary et al., 2010; Waszczuk et al., 2010).

Information obtained from this step 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 (Recasens et al., 2010) format, persisting some of the morphosyntactical information and mention boundaries, as well as mention head indication, where applicable.

2.2. Coreference resolution

The input to the coreference resolution module consists of two corpora stored in SemEval format. One corpus acts as the gold standard and the other one is (as a result of rule-based classification) automatically filled with mention chains information. These two files in the end are compared using script provided by SemEval organisers, calculating a number of different evaluation measures.

The implemented coreference resolution module uses a standard best-first entity-based model based on syntactic constraints (elimination of nested nominal groups), syntactic filters (elimination of syntactic incompatible heads), semantic filters (wordnet-derived compatibility) and selection (weighted scoring). As was mentioned before, morphosyntactic properties are obtained from Spejd, Pantera and NER. Semantic properties are currently based on Polish WordNet (Piasecki et al., 2009).

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 distance measure defining the term ”closest” takes into consideration not only the word distance of the last word of the mention but also the depth of nesting. This means that if two mentions are nested, the inside one is referred to be further and therefore in case of a tie it is not going to be chosen.

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 the 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 following five rich rules:

1. *gender/number rule* eliminates syntactically incompatible matches, i.e. marks mentions with different gender or number as incompatible,
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,
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).

3. Data sets and evaluation

Evaluation data came from the balanced part of the National Corpus of Polish (Przepiórkowski et al., 2008) which provided 50 randomly selected text samples (20 sentences each) containing altogether about 6500 mentions (≈ 1000 sentences, ≈ 20000 tokens). 35 texts were used as development data and 15 texts with 1737 mentions as testing data. Average mention had the length of 1.9 tokens.

For evaluation, the test data files have been automatically preprocessed 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.

Among 1737 mentions in gold standard data 1262 mention chains were formed. 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 Table 1.

The implemented system was designed to provide an environment for testing coreference rule sets, which facilitated creating two common variations: the all-singletons and head-match baselines plus slightly more

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

Table 1. Mention chain length

System type	**MUC**			**CEAFE**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons	–			93.10%	67.64%	78.35%
All-singletons + head match	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-singletons	72.65%	100.00%	84.16%	50.00%	49.18%	49.58%
All-singletons + head match	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. Results of the experiment with gold standard mentions

complex, although still very straightforward 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.

Another capability of the system is to test end-to-end system performance (meaning raw text as input and mention chains as an output) as well as the version of the coreference resolution task with already tagged gold standard mention boundaries as input data. 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.

4. Experimental results

The formal experimental results are presented in Tables 2–4 for all the recognized metrics: MUC (Vilain et al, 1995), CEAF (Luo, 2005), B^3 (Bagga and Baldwin, 1998) and BLANC (Recasens and Hovy, 2010).

Table 2 shows the results for the system running on gold standard mention boundaries. Table 3 presents the scores for the fully-fledged end-to-end system taking raw text as input data (with mention detection carried out by a submodule). Table 4 presents the performance of the end-to-end system in a variant neglecting the zero anaphora phenomenon.

The reason for treating the last case separately is lack of the zero anaphora detection module in the current version of the system. As the results prove, zero anaphora a really important factor in the evaluation of coreference resolution system – the impact of removing it from evaluation makes a huge impact on the scores. Therefore, a high-quality zero anaphora detector would provide a significant improvement to performance of the end-to-end system. 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 (e.g. as with „*land*" and „*part*", resulting from having „*native land*" and „*parts*" in „*coming from the same parts*").

4.1. Mention-detection

Results of automatical mention-detection are as follows: system achieved recall of 83.82% and precision of 78.71%, which results in F1 measure score of 81.18%. However, as it lacks zero anaphora detection submodule, it's also worth to report the score without zero anaphora. In this case, recall rises to 88.86%, and as precision remains the same, final F1 score is 83.48%.

4.2. Towards manual error classification

Another way of assessing the accuracy of coreference resolution system is based on the structure of mention chains. Each mention in a chain (except of course the last chain element) has exactly one mention being the next part of that chain. It means that coreference itself can be understood as the link between a mention and next mention (which refers to the same entity) in the discourse. Such understanding allows a manual assessment of results, because it brings the computational complexity of other measures to a manageable level of simply comparing one link at a time, processing the text in a linear manner. Gold standard annotation of mention chains in the test corpora has 1067 mentions, which do not have a subsequent mention in their mention chain. In 475 cases a mention has a link to following mention, pointing to the same entity.

System type	MUC			CEAFE		
	R	P	F1	R	P	F1
All-singletons	–			44.04%	29.94%	35.65%
All-singletons + head match	16.63%	16.80%	16.71%	38.36%	34.93%	36.56%
5 rules	17.26%	14.04%	15.48%	35.88%	35.60%	35.74%
4 rules (no wordnet)	17.26%	15.53%	16.35%	37.65%	35.78%	36.69%
	B^3			BLANC		
	R	P	F1	R	P	F1
All-singletons	32.61%	40.46%	36.11%	50.00%	29.05%	36.75%
All-singletons + head match	35.61%	33.05%	34.28%	50.33%	59.24%	37.99%
5 rules	35.74%	30.30%	32.80%	50.27%	55.37%	38.18%
4 rules (no wordnet)	35.74%	31.98%	33.75%	50.35%	58.57%	38.13%

Table 3. End-to-end experiment with zero anaphora

System type	MUC			CEAFE		
	R	P	F1	R	P	F1
All-singletons	–			85.93%	58.15%	69.36%
All-singletons + head match	58.24%	48.08%	52.68%	76.61%	69.42%	72.84%
5 rules	65.20%	43.32%	52.05%	71.49%	70.59%	71.03%
4 rules (no wordnet)	64.43%	47.34%	54.58%	75.70%	71.60%	73.59%
	B^3			BLANC		
	R	P	F1	R	P	F1
All-singletons	69.58%	80.92%	74.82%	50.00%	46.45%	48.16%
All-singletons + head match	81.15%	71.14%	75.81%	53.95%	79.34%	55.54%
5 rules	82.64%	65.91%	73.33%	54.20%	72.48%	55.86%
4 rules (no wordnet)	82.42%	69.24%	75.26%	54.26%	77.60%	56.03%

Table 4. End-to-end experiment without zero anaphora

The best system described in this paper is the one using 4 rules in the gold mention boundaries task. It managed to correctly classify 873 of 1067 mentions (≈ 82%) as the ones having no link to next mention, in 303 cases out of 475 (≈ 64%) it found a correct connection. The extension of this simple error classification would consist of manual assessment of each of this examples and clustering them into a number of various groups, discriminated based on the reason of the failure of the system.

4.3. Manual mention discrimination

The useful side-effect of the creation and evaluation of described system was the identification of some decisions required to create a well-defined guide for linguists doing manual mention detection for the purpose to create gold standard corpus.

First important choice was to tag mentions consisting of subsequent words only. This decision was made for simplicity purposes, but a more complex system supposedly should consider the possibility of having non-continuous mentions. SemEval format doesn't allow to tag such mentions, probably because there are not many (if any) cases for such phenomenon to appear for English, but texts in Polish language with its free word order can contain non-continuous mentions.

Another interesting decision was the choice of additional information to be annotated with a mention tag. In the current version of the system this information contains the mention type (one of: noun, pronoun, named entity) which is used in the pronoun rule, affecting pronouns only. As this information is valid for individual rules, more well-grained distinctions can be crafted.

5. Conclusions and further work

The presented approach is a first experimental step towards general-purpose coreference resolution for Polish. It builds on latest findings in the field, the most important of which being the precedence of few rich features over the multitude of weak ones. 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 (Woliński, 2004) as well as information extracted from Polish Wikipedia and other available fact bases). Another very useful improvement would be to create a zero anaphora detector for Polish and integrate it into the end-to-end version of experiment, as zero anaphora was discovered to contribute very significantly to evaluation measures.

The results of this process are also intended to create synergy with ATLAS project[1] where anaphora

[1] Applied Technology for Language-Aided CMS co-funded by the European Commission under the

resolution module is planned to be integrated in the summarization component.

6. References

Acedański, S. and Gołuchowski, K. (2009). A Morphosyntactic Rule-Based Brill Tagger for Polish. In *Recent Advances in Intelligent Information Systems*, pp. 67–76, Kraków, Poland. Academic Publishing House EXIT.

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

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In: *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pp. 563–566.

Haghighi, A. and Klein, D. (2009). Simple coreference resolution with rich syntactic and semantic features. In: *EMNLP'09*, pp. 1152–1161.

Luo, X. (2005). On Coreference Resolution Performance Metrics. In: *Proceedings of HLT-EMNLP*, Vancouver, Canada, pp. 25–32.

Vilain, M., Burger, J. Aberdeen, J., Connolly, J. and Hirschman, L. (1995). A Model-Theoretic Coreference Scoring Scheme. In: *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52.

Marciniak, M. (2002). Anaphor Binding in Polish. Theory and Implementation. In: *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, Lisbon, Portugal.

Matysiak, I. (2007). Information Extraction Systems and Nominal Anaphora Analysis Needs. In: *Proceedings of the International Multiconference on Computer Science and Information Technology*, pp. 183–192, Wisła, Poland.

Mitkov, R., Belguith, L. and Styś, M. (1998). 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.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In: *Proceedings of the ACL*, pp. 1396–1411, Uppsala, Sweden.

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

Piasecki, M., Szpakowicz, S. and Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wrocławskiej.

Przepiórkowski, A. and Buczyński, A. (2007). Spejd: Shallow parsing and disambiguation engine. In: *Proceedings of the 3rd Language & Technology Conference*, Poznań.

Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467).

Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B. and Łaziński, M. (2008). Towards the National Corpus of Polish. In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008*), Marrakech, Morocco. ELRA.

Recasens, M. Marquez, L., Taulé, M. Hoste, V, Poesio, M., Martí, M. A. and Versley, Y. (2010). *SemEval-2010 Task 1: Coreference Resolution in Multiple Languages,* pp. 70–75. Association for Computational Linguistics.

Recasens, M. and Hovy, E. (2010). BLANC: Implementing the Rand index for coreference evaluation. In: *Natural Language Engineering*, pp. 1–26.

Savary, A., Waszczuk, J. and Przepiórkowski, A. (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.

Waszczuk, J., Głowińska, K., Savary, A. and Przepiórkowski, A. (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)*, pp. 531–539, Wisła, Poland. PTI.

Woliński, M. (2004). *Computer-aided verification of Świdziński's grammar*. PhD dissertation, Warsaw. [In Polish]. Institute of Computer Science, Polish Academy of Sciences.

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