Rule-based coreference resolution for Polish

Maciej Ogrodniczuk and Mateusz Kopeć
Institute of Computer Science, Polish Academy of Sciences

CORE project

General information

The *Computer-based methods for coreference resolution in Polish texts* project (CORE) financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40).

Project time frame: 2011–2014.

Project mission

Create methods and tools for **automated anaphora and coreference resolution of Polish** by preparation of:

- Typology of Polish coreference.
- Polish coreferential corpus – a subset of the National Corpus of Polish (NKJP) manually annotated with coreferential chains.
- IT tools for coreference resolution (rule-based, statistical, hybrid) and their evaluation.

SYSTEM DESCRIPTION

Scope of the current task

Create **the first noun phrase coreference resolution** system for Polish, intended to provide the starting ground for further experiments and a useful reference baseline.

Mention detection

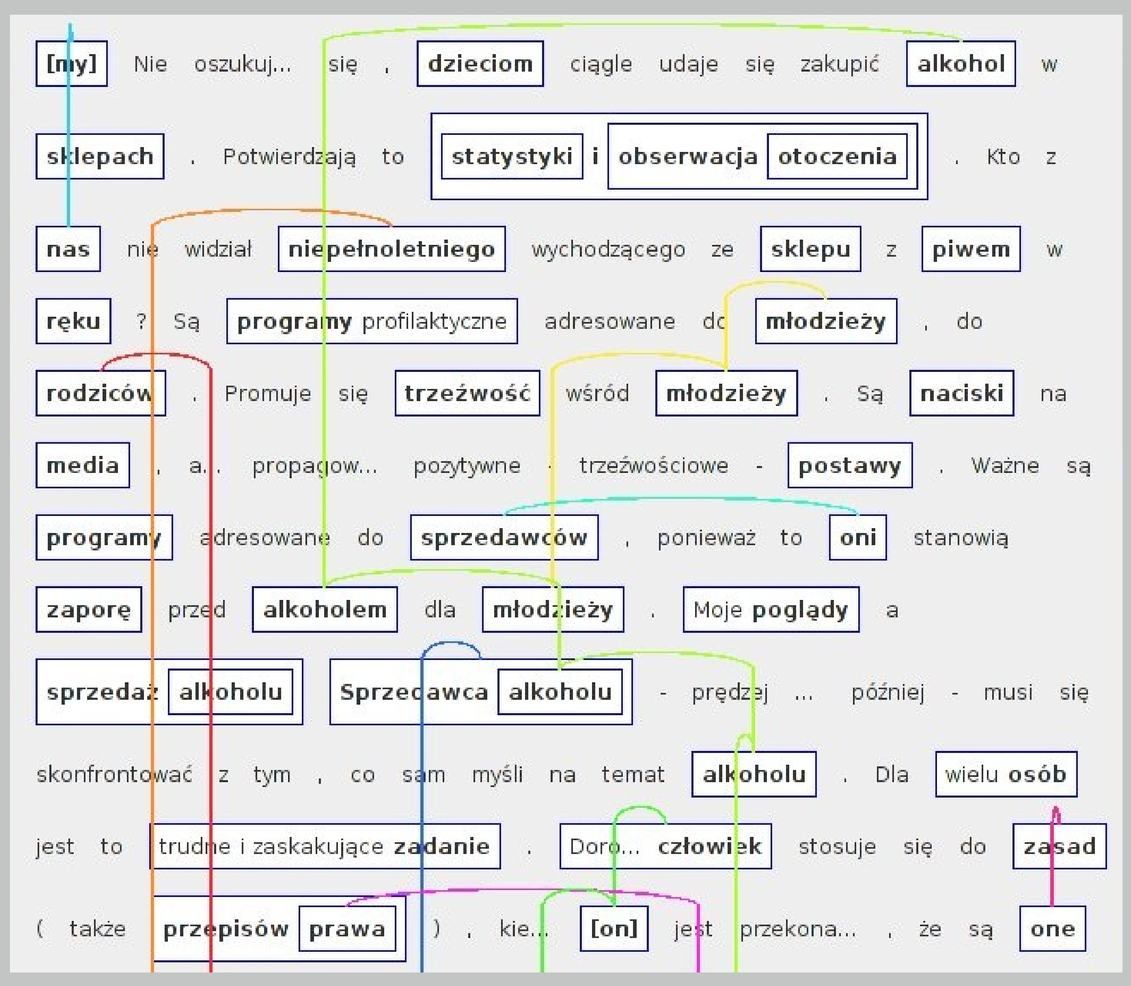
- **POS tagging** with Pantera/Morfeusz SGJP,
- **NP chunking** with Spejd shallow parser,
- **NE recognition** with NER tool,
- **Wordnet-based** processing with plWordNet.

Coreference resolution

Few rich linguistic features (cf. Haghighi and Klein):

1. **syntactic constraints** (elimination of nested nominal groups),
2. **syntactic filters** (elimination of syntactic incompatible heads),
3. **semantic filters** (wordnet-derived compatibility),
4. **selection** (weighted scoring).

Visualisation: internal prototype



DATA SETS AND EVALUATION

Rule set

1. **gender/number rule** eliminates syntactically incompatible matches (e.g. wrt. gender or number),
2. **including rule** eliminates nested groups,
3. **lemma rule**, for nominal groups only, promotes head matches,
4. **wordnet rule**, for nominal groups with wordnet representation; investigates synonyms, hyperonyms, alternyms and fuzzynyms,
5. **pronoun rule**, promotes matching pronouns.

Working data set

- 50 randomly selected text samples from NKJP,
- 20-sentence-length,
- 6500 mentions altogether,
- average mention length: 1.9 tokens,

Evaluation data

- 15 randomly selected text samples,
- 1737 mentions,
- 1262 mention chains,
- average chain size: 1.37 mentions.

Chain length	1	2	3	4	5	6	7..10	11..27
Number of chains	1079	88	43	20	9	6	2..5	1

Experimental results

System type	**MUC**			**CEAF**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons		–		85.9%	58.2%	69.4%
All-singletons + head	58.2%	48.1%	52.7%	76.6%	69.4%	72.8%
5 rules	65.2%	43.3%	52.1%	71.5%	70.6%	71.0%
4 rules (no wordnet)	64.4%	47.3%	54.6%	75.7%	71.6%	73.6%
	B^3			**BLANC**		
	R	**P**	**F1**	**R**	**P**	**F1**
All-singletons	69.6%	80.9%	74.8%	50.0%	46.5%	48.2%
All-singletons + head	81.2%	71.1%	75.8%	54.0%	79.3%	55.5%
5 rules	82.6%	65.9%	73.3%	54.2%	72.5%	55.9%
4 rules (no wordnet)	82.4%	69.2%	75.3%	54.3%	77.6%	56.0%

CONCLUSIONS

Next steps

- **zero anaphora detection** experiments,
- wider range of coreference constructs such as **identity of sense**,
- **typization** of coreferential links,
- **refinement of grammar** used for identification of mentions,
- **machine learning experiments**,
- **feature base expansion** (from deep parse results, fact bases etc.)

Synergies with CIP ICT-PSP projects

- ATLAS – Applied Technology for Language-Aided CMS (`http://www.atlasproject.eu`): CR for text summarization,
- CESAR – CEntral and South-east europeAn Resources, part of META-NET (`http://www.meta-net.eu/projects/cesar`) – Polish LRTs made available in META-SHARE repository.