End-to-end coreference resolution baseline system for Polish

Maciej Ogrodniczuk | Institute of Computer Science
Mateusz Kopeć | Polish Academy of Sciences

CORE

5th Language and Technology Conference
Poznań, November 26, 2011

The CORE project

Project factsheet

- Computer-based methods for coreference resolution in Polish texts
- A National Science Centre grant 6505/B/T02/2011/40
- Duration: 2011-2014
- Principal investigator: Maciej Ogrodniczuk

Project summary

1. Create innovative methods and tools for automated anaphora and coreference resolution in Polish texts
2. Create a corpus of Polish annotated with coreferential chains
3. Test various coreference resolution approaches on the annotated data (rule-based, statistical, hybrid etc.)

Noun phrase coreference resolution

Task definition

1. NP = a group of adjacent words having nominal head, e.g. pronouns, proper nouns, nominal groups etc.
2. Nesting allowed (*dyrektor departamentu* = EN: *director of the department*)
3. *Identity of reference* only

Two-step process

1. Identify mentions
2. Build coreference chains with mentions having identical referent

Mention detection

3 steps

1. POS tagging with PANTERA
2. NP chunking with SPEJD shallow parser
3. NE recognition with NER tool

Heuristics

1. Elimination of mentions with the same boundaries
2. Elimination of mentions with the same head
3. Preference of longer mentions

Coreference resolution

CORE

Resolution algorithm

```
for each mention (in order of appearance):
  from mention chains (already found):
  find the chain with maximal similarity(mention,chain)
  if similarity(mention,chain) > threshold:
    add the mention to the chain
  else:
    create a new mention chain with the mention
```

Similarity calculation

1. similarity(m,ch) $= max_{n \in ch}$ similarity(m, n)
2. similarity between mentions is calculated by applying a set of rules

Rule set

Rules

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

Tie-breaker

1. choose the closest mention
2. (including nesting)

Evaluation data

Data statistics

- from the balanced part of the National Corpus of Polish
- 15 texts of 20 sentences
- 1737 mentions, average mention length: 1.9 tokens
- 1262 mention chains, average mentions in chain: 1.37

Mention chain size

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

Experimental results

Mention detection

- With zero anaphora: R: 83.82%, P: 78.71%, F1: 81.18%
- Without zero anaphora: R: 88.86%, P: 78.71%, F1: 83.48%

Coreference resolution

Four rule sets:

1. All-singletons
2. All-singletons + head match
3. 5 rules
4. 4 rules (no wordnet)

End-to-end with zero anaphora

System type	MUC			CEAF		
	R	P	F1	R	P	F1
All-singletons	–			44.04%	29.94%	35.65%
All-singl. + head m.	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-singl. + head m.	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%

End-to-end without zero anaphora

System type	MUC			CEAF		
	R	P	F1	R	P	F1
All-singletons	–			**85.93%**	58.15%	69.36%
All-singl. + head m.	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-singl. + head m.	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%**

Results with gold standard mentions

System type	MUC			CEAF		
	R	P	F1	R	P	F1
All-singletons	–			**93.10%**	67.64%	78.35%
All-singl. + 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-singletons	72.65%	**100.00%**	84.16%	50.00%	49.18%	49.58%
All-singl. + 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%**

Conclusions

CORE

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 – `www.atlasproject.eu`: CR for text summarization
- CESAR – `www.meta-net.eu/projects/cesar`: Polish LRTs made available in META-SHARE repository

Newest findings

Adapting foreign CR systems for Polish:

1. rule-based approaches:
   RARE – Robust Anaphora Resolution by University of Iasi,
2. statistical approaches:
   BART – Baltimore/Beautiful Anaphora Resolution Toolkit.

Rethinking the notion of identity:

1. Identity vs. near-identity
2. NIDENT typology (Recasens and Hovy)
3. Refocusing and neutralization

Thank you!

CORE

It's question time!