Nested Mention Detection for Polish Coreference Resolution

Maciej Ogrodniczuk[1], Alicja Wójcicka[1], Katarzyna Głowińska[1,2], Mateusz Kopeć[1]

[1] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland

[2] Lingventa, Reymonta 10/25, 01-842 Warsaw, Poland

PCC Mention Model vs. NKJP Grammar

NKJP Grammar in CORE project

- NKJP Grammar is a shallow grammar of Polish used by Spejd parser to provide syntactic annotation of the 1-bilion-word part of the National Corpus of Polish.
- It was used by CORE project for annotation of mentions - nominal groups referencing discourse-world objects in Polish Coreference Corpus.
- New rules have been incorporated into grammar in order to detect nested mentions, e.g. *the CEO of Microsoft*.

Mentions in PCC

- Mentions in PCC are all nominal phrases (NGs) — syntactic groups with nominal or pronominal heads (syntactic and/or semantic).
- A nested nominal phrase is marked as separate from the superior phrase when its syntactic/semantic head is other than the head of the superior phrase.
- The PCC nominal phrase consists of:
  - adjectives
  - nouns
  - gerunds
  - conjunctions (coordinated groups)
  - subordinate numerals
  - superordinate numerals
  - relative subordinate clauses
  - prepositional phrases,
  - adjectival participles.
- All potentially referential constructs are marked.

Nominal groups in the NKJP project

- Syntactic annotation in the National Corpus of Polish was limited to joining words together into constituents.
- The nominal groups in the NKJP project were extensive — they consisted of as many elements as possible, for e.g. in a phrase composed of consecutive nouns in the genitive case such as *propozycji wyznaczenia daty rozpoczęcia procesu wprowadzania reformy ustroju*. *'proposal for setting the date of launching the process of introducing reform of the system'*, the whole phrase was the only detected nominal group.

Mention Detection Chain

Preprocessing

The processing of a raw text begins with part-of-speech tagging with Pantera. Then the text is shallow parsed with Spejd and its morphological component Morfeusz SGJP. The last step is to detect Named Entities, which is done by NER. Information obtained from this step is then used to collect mention boundaries

Mention Detection Process

MentionDetector works in three steps:

1. It collects mention candidates from morphosyntactic, shallow parsing and/or named entity level (lack of any layer simply results in fewer mention candidates discovered) and also produces zero-anaphora candidates.
2. It removes redundant/unnecessary candidates.
3. It updates head information among mentions.

Towards the New Grammar

Rule Modification

The structure of the section of rules detecting syntactical groups was modified.

- Rules for syntactic groups without nesting are in the new version of the grammar separated from rules for groups with nesting and are placed before them.
- Groups without nesting should contain only syntactic words, even if they have complicated structure, containing e.g. adjectives and particles or numerals (as in a group: *kilka kolejnych filii szkolnych 'a few other school branches'*).
- The most problematic are rules detecting nominal-nominal groups without nesting, e.g. proper names of persons (*Jan Kowalski*) or appositions (*malarz pejzażysta 'landscape painter'*).
- The part of the grammar responsible for nested groups is built in another manner. The only elements of these groups are other syntactical groups, nested or not nested.

Nested Groups

- There are four main types of nested groups: case-governed groups, prepositional groups, coordinated groups (conjunction governed groups) and relative clauses.
- Different types of groups with nesting can be embedded in all other types of groups.
- The order of the rules is as follows:
  1. the group of rules detecting case-governed groups, restricted only to the context without comma or conjuction on the right side of the given string
  2. the rules responsible for coordinated groups
  3. the rules detecting case-governed groups, without the restriction mentioned above

Evaluation

Table: Evaluation results, setting 1

		NKJP Grammar	New version
Mention statistics	**Total gold mentions**	53,407	53,407
	Total system mentions	51,217	51,750
	Total common mentions	33,839	34,176
Mention detection results	**Precision**	66.07%	66.04%
	Recall	63.36%	63.99%
	F1	64.69%	65.00%

Table: Evaluation results, setting 2

		NKJP Grammar	New version
Mention statistics	**Total gold mentions**	53,407	53,407
	Total system mentions	65,853	69,475
	Total common mentions	31,582	33,122
Mention detection results	**Precision**	47.96%	47.67%
	Recall	59.13%	62.02%
	F1	52.96%	53.91%

Acknowledgements

- The work reported was carried out within the "Computer-based methods for coreference resolution in Polish texts" (CORE) project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40). The work was also co-funded by the European Union from the resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications".

http://clip.ipipan.waw.pl/CORE/ m.ogrodniczuk@ipipan.waw.pl, alicja.wojcicka@wp.pl, m.kopec@ipipan.waw.pl, k.glowinska@gmail.com