LRT for Polish coreference annotation and resolution

Mateusz Kopeć and Maciej Ogrodniczuk

mateusz.kopec@ipipan.waw.pl, maciej.ogrodniczuk@ipipan.waw.pl

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5
01-248 Warsaw, Poland

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.

ANNOTATION PROCESS

Process description

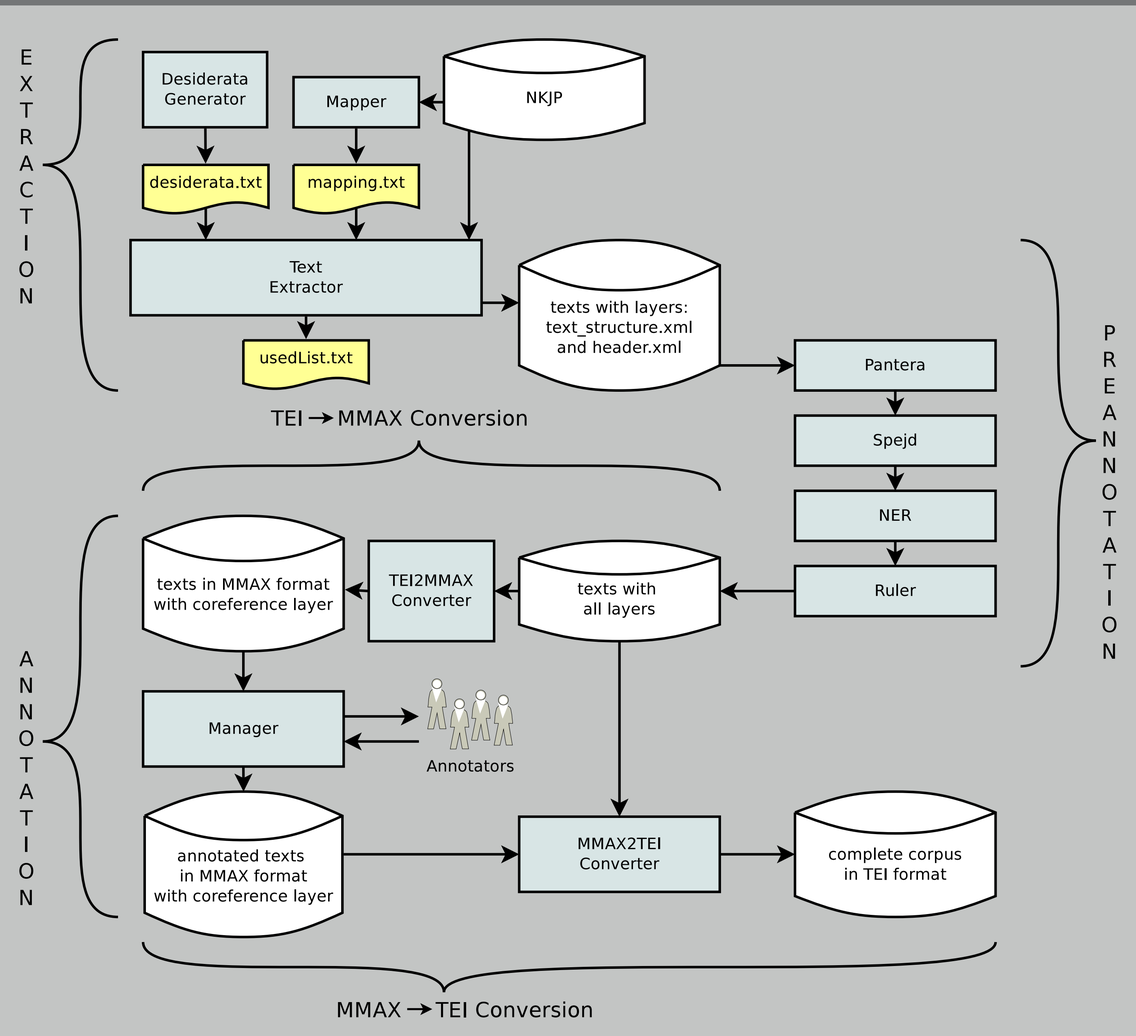

Figure: Process organization

Preprocessing

1. POS tagging with Pantera/Morfeusz SGJP (`http://clip.ipipan.waw.pl/PANTERA`),
2. NP chunking with Spejd shallow parser (`http://clip.ipipan.waw.pl/Spejd`),
3. NE recognition with NERF tool (`http://clip.ipipan.waw.pl/Nerf`),
4. Mention detection and coreference resolution by RULER (`http://clip.ipipan.waw.pl/RULER`).

Manual annotation tools

Annotators use two tools:

- Manager (client) – a program for acquiring texts from the server and sending them back when annotated (also used in other project),
- Mmax – a tool for single text annotation (based on MMAX2 tool by Müller and Strube).

Annotation scope

Annotators are supposed to mark in each text:

- mentions,
- mentions' semantic heads,
- clusters of coreferent mentions,
- dominant phrase of each mention cluster,
- quasi-identity links.

Each text (almost, see the "Agreement" sect.) is:

- annotated by one annotator,
- superannotated (checked and corrected) by one superannotator.

Annotation guidelines

What is a mention?

- A noun group (NG) – noun, possibly accompanied by modifiers, personal pronouns, etc. (marked with as wide borders as possible)...
- ...except some cases:
  - reflexive pronouns (*się "myself"*),
  - reciprocal pronouns (*siebie "each other"*),
  - demonstrative pronouns introducing subordinates other than relative clauses (*o tym, że "of-this-that = of the fact that"*),
  - interrogative pronouns (*kto "who"*),
  - indefinite pronouns (*ktoś "somebody"*),
  - negative pronouns (*nic "nothing"*),
  - possessive pronouns, which behave like adjectives in Polish (*mój "mine"*).
- Zero subject (marked by annotators at the verb).

What are quasi-identity links? They connect two mentions:

- either suggested as by the text as identical but not identical in reality (*Wziął wino z lodówki i wypił je. "He took the wine from the fridge and drank it."*),
- or suggested as by the text as not identical but identical in reality (*Warszawa przedwojenna i ta z początku XXI wieku "Prewar Warsaw and the one at the beginning of the 21st century"*).

THE CORPUS

Corpus architecture

Texts are samples of size 250-350 segments each, extracted randomly from the National Corpus of Polish. Text type proportions follow these used in NKJP.

Texts type	# of texts	# of segments[1]	Percent[2]
Dailies	459	127500	25.5%
Magazines	406	117500	23.5%
Fiction literature (prose, poetry, drama)	288	80000	16%
Non-fiction literature	96	27500	5.5%
Instructive writing and textbooks	100	27500	5.5%
Spoken – conversational	83	25000	5%
Internet – interactive (blogs, forums, usenet)	63	17500	3.5%
Internet – non-interactive (static pages, Wikipedia)	63	17500	3.5%
Miscellaneous written (legal, advertisements, user manuals, letters)	55	15000	3%
Spoken from the media	44	12500	2.5%
Quasi-spoken (parliamentary transcripts)	43	12500	2.5%
Academic writing and textbooks	35	10000	2%
Unclassified written	19	5000	1%
Journalistic books	19	5000	1%
Total	1773	500000	100%

Table: Corpus text types balance

[1] at least – but no more that 500 segment difference in each row (beside total)
[2] approximately

Current annotation status – 17.09.2012

- 1453/1773 texts (413519/504000 segments) – annotated (82%):
  - 136565 mentions, 84382 singletons (61.8%),
  - 14129 clusters,
  - 3582 near-identity links.

Mention size	1	2	3	4	5	6	7	8	9	10	...	235	Any
# mentions	67634	30631	13020	7192	4728	3103	2140	1628	1219	982	...	1	136565

Table: Mentions sizes in segments

Mention cluster size	1	2	3	4	5	6	7	8	9	10	11	...	41	Any >1
# clusters	84382	7636	2524	1213	739	476	317	258	165	140	105	...	1	14129

Table: Mention clusters size statistics

- 424 texts out of 1773 – superannotated (24%).

Agreement between annotators

Part of the corpus – 210 texts (60674 segments), taken equally from each type (15 texts each) was annotated by two annotators independently to check the agreement.

The results are as follows:

- mentions agreement: $F_1 = 85.55\%$ (based on normal precision/recall of full mentions in both annotations),
- mentions' semantic heads $\kappa = 97\%$ (with adjustment for chance agreement with uniform head choice probability distribution),
- clusters of coreferent mentions:
  - $\kappa = 74.24\%$ (agreement of decision: "singleton"/"in cluster" for each mention, with adjustment for chance agreement with probability distribution calculated from all texts),
  - $\kappa = 77.5\%$ (agreement of coreference and non-coreference links as in BLANC measure, with adjustment for chance agreement with probability distribution calculated from particular text),
- dominant phrase of each mention cluster: $acc = 63,04\%$,
- near-identity links: $\kappa = 22.2\%$ (with adjustment for chance agreement with probability distribution calculated from particular text).

AUTOMATIC ANNOTATION TOOLS

RULER

- A rule-based baseline coreference resolver and mention detector:
  - detects mentions using data from other preprocessing tools,
  - clusters mentions into coreference groups,
  - doesn't detect quasi-identity.

BART

- Well-known machine learning multilingual coreference resolver:
  - first experiments of adapting to Polish already conducted,
  - doesn't provide superior performance off-the-shelf, but needs further tweaking.

FUTURE

Next steps

- Finalize the corpus annotation and superannotation (by the end of 2012).
- Find and eliminate annotation errors (as they allows ooccur in human annotation).
- Analyze quasi-identity annotation data.
- Evaluate existing CR systems on the corpus (for example RULER, BART).
- Develop tools designed for Polish:
  - coreference resolution,
  - mention resolution and zero-subject detection.
- Use coreference resolution for the benefit of other language processing tasks, including:
  - summarization,
  - text categorization.