# Creating a Coreference Resolution System for Polish

Mateusz Kopeć and Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 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.

## SYSTEM DESCRIPTION

### Scope of the current task

Adapt a well-known statistical system – BART: Beautiful Anaphora Resolution Toolkit (Versley et al., 2008) to Polish language and initially compare it with the first, rule-based approach and provide valuable experience for the multilingual users of BART.

### BART architecture

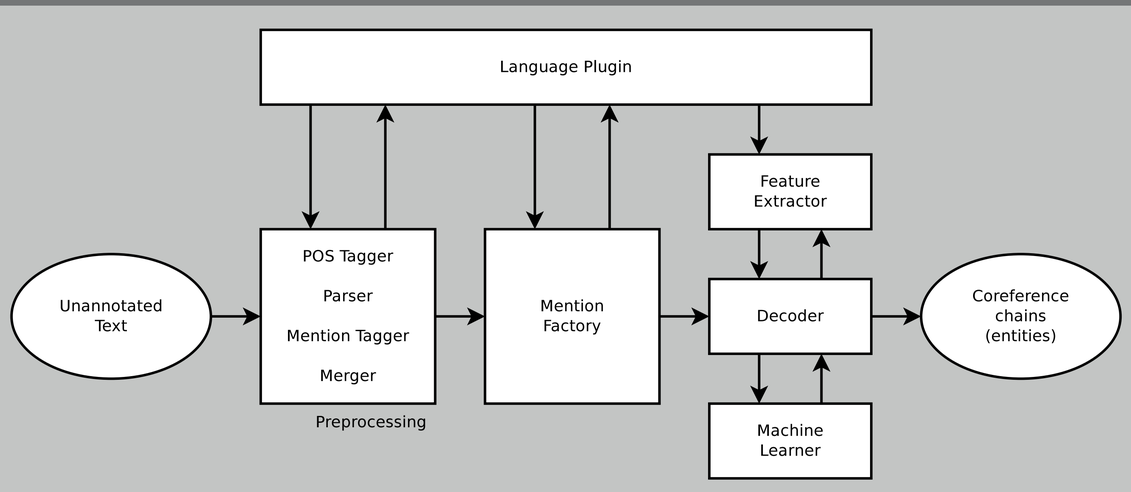


Figure: Language-agnostic architecture of BART

### BART for Polish

Adjusting BART for Polish required following steps:

- preprocess the coreference corpus to add morphosyntactic, shallow parse and named entity layers,
- convert the corpus to MMAX format (Müller and Strube, 2006) with 3 layers:
  - the segmentation layer,
  - the markable layer,
  - the coreference layer,
- select language-agnostic or language-adaptable feature extractors,
- configure the Polish Language Plugin,
- conduct the experiment.

### Preprocessing

Preprocessing was carried out outside BART and involved:

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`).

### Used features

- `First_Mention` – extracting information, whether given mention is the first one in its mention chain
- `FirstSecondPerson` – checking if mentions are first or second person
- `Gender`, `Number` – extracting compatibility of gender/number of two mentions
- `HeadMatch` – comparing heads of mentions
- `MentionType`, `MentionType_Anaphor`, `MentionType_Salience` – providing a number of features based on mention types (for example if they are pronouns or reflexive pronouns)
- `DistDiscrete`, `SentenceDistance` – providing information about text distance between mentions in terms of sentences
- `StringKernel`, `StringMatch`, `LeftRightMatch` – feature extractors based on orthographic similarity of mentions.

## DATA SETS AND EVALUATION

### Evaluation data

- 15 randomly selected text samples of about 20 sentences each,
- 1737 mentions,
- 1262 mention chains,
- the average size of mention chain: 1.37 mentions.

| **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: Mention chains size statistics

### Experimental results

| **System type** | **MUC** | | |
|---|---|---|---|
| | **R** | **P** | **F1** |
| BART | 65.11% | 58.06% | 61.38% |
| Rule-based | **66.23%** | **63.77%** | **64.98%** |
| | **$B^3$** | | |
| | **R** | **P** | **F1** |
| BART | **89.17%** | 87.27% | 88.21% |
| Rule-based | 88.94% | **89.81%** | **89.37%** |
| | **CEAFM** | | |
| | **R** | **P** | **F1** |
| BART | 82.34% | 82.34% | 82.34% |
| Rule-based | **83.94%** | **83.94%** | **83.94%** |
| | **CEAFE** | | |
| | **R** | **P** | **F1** |
| BART | 83.80% | 87.06% | 85.40% |
| Rule-based | **86.54%** | **87.59%** | **87.06%** |
| | **BLANC** | | |
| | **R** | **P** | **F1** |
| BART | **76.20%** | 81.09% | 78.43% |
| Rule-based | 75.10% | **83.70%** | **78.75%** |

Table: Comparison of two systems

## CONCLUSIONS

### Next steps

- Train and evaluate BART on a bigger corpus of better quality, which annotation is under way,
- adapt for Polish other machine learning coreference resolution tools such as RARE (Cristea et al., 2002) and Reconcile (Stoyanov et al., 2010) and compare their accuracy with BART,
- incorporate existing preprocessing tools for Polish into BART.