Creating a Coreference Resolution System for Polish

Mateusz Kopeć, Maciej Ogrodniczuk

Institute of Computer Science
Polish Academy of Sciences
ul. Jana Kazimierza 5, Warsaw, Poland
mateusz.kopec@ipipan.waw.pl, maciej.ogrodniczuk@ipipan.waw.pl

Abstract

Although the availability of the natural language processing tools and the development of metrics to evaluate them increases, there is a certain gap to fill in that field for the less-resourced languages, such as Polish. Therefore the projects which are designed to extend the existing tools for diverse languages are the best starting point for making these languages more and more covered. This paper presents the results of the first attempt of the coreference resolution for Polish using statistical methods. It presents the conclusions from the process of adapting the Beautiful Anaphora Resolution Toolkit (BART; a system primarily designed for the English language) for Polish and collates its evaluation results with those of the previously implemented rule-based system. Finally, we describe our plans for the future usage of the tool and highlight the upcoming research to be conducted, such as the experiments of a larger scale and the comparison with other machine learning tools.

Keywords: coreference resolution, BART, anaphora resolution, machine learning

1. Introduction

The statistical methods are well-known to be very successful for many natural language processing tasks, including the coreference resolution. Nevertheless such attempt has so far never been made for Polish, mostly because of lack of the coreference annotation methodology and the evaluation data. The process targeted at changing this situation has already been started with the *Computer-based methods for coreference resolution in Polish texts* project which aims at creating the coreferential corpus of Polish manually annotated with various types of identity of reference with near-identity relations, similarly to (Recasens et al., 2010a)).

First experiments on the rule-based coreference resolution of Polish (Ogrodniczuk and Kopeć, 2011a; Ogrodniczuk and Kopeć, 2011b), apart from Mitkov et al.'s work on multilingual anaphora resolution which also included Polish (Mitkov et al., 1998), have already shown their usefulness in gathering experience for the next phases of the project and resulted in creating the first set of Polish data manually annotated with mentions and coreferential chains. The present attempt at using a well-known statistical system – BART: Beautiful Anaphora Resolution Toolkit (Versley et al., 2008) – allows to initially compare these two approaches and provides valuable experience for the multilingual users of BART.

2. BART and the Polish Language Plugin

Beautiful Anaphora Resolution Toolkit is a system for performing automatic coreference resolution, including necessary preprocessing steps. It allows to test various machine learning approaches, such as the algorithms from Weka (Witten et al., 1999) or the Maximum Entropy model (Berger et al., 1996). As an open-source tool with a modular design it proves to be easily adaptable for languages other than English to create a statistical baseline system for coreference resolution.

BART's modularity (see Fig. 1[1]) involves separation of two tasks: the preprocessing of texts, resulting in mention detection, and the automatic coreference resolution, understood as a machine learning task. As preprocessing tools included in the toolkit are designed specifically for English, preprocessing for the Polish texts for the experiments was carried out outside BART.

The machine learning approach requires training examples to be annotated with features and mention chains. BART offers 64 feature extractors to transform the training examples into features, however using them out-of-the-box for languages other than English is problematic due to their language-specific settings. Although some of them are extracted into the *Language Plugins*, which are supposed to increase the modularity of the toolkit by discriminating the non-language-agnostic parts of BART, a large number of the feature extractors still contain the settings specific for English. For example, a feature extractor may take into consideration a specific (English) substring of the mention or the English definite article, not to mention obvious cross-lingual tagset incompatibilities. Another difficulty, this time objective, arises from the lack of certain types of language processing tools for Polish. Taking these into account, only 13 pair feature extractors were selected for the experiments:

- `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

The work reported here 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).

[1] Cf. Example system configuration in (Versley et al., 2008), Fig. 2.

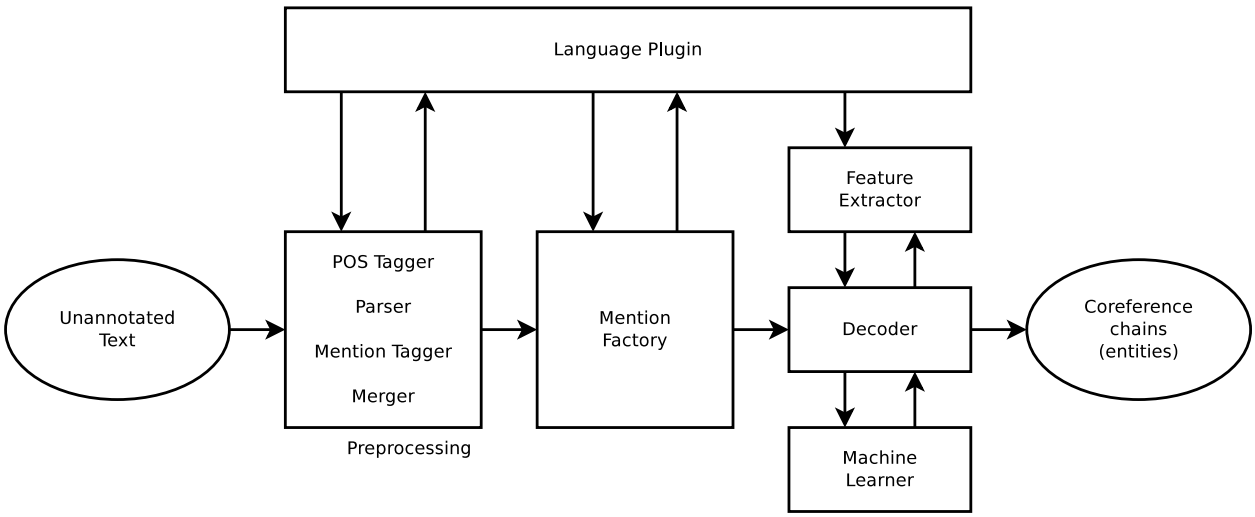

Figure 1: BART architecture

- `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`, `LeftRight-Match` – feature extractors based on orthographic similarity of mentions.

For the purpose of described experiments the *Polish Language Plugin* has been implemented to transform tagset and morphological information into BART features. It was based on similar plugins available for German and Italian.

3. Data Set and Evaluation

Data Source The texts for BART experiments have been extracted from the National Corpus of Polish (Przepiórkowski et al., 2008) and automatically pre-processed with the noun phrase chunker Spejd (Przepiórkowski and Buczyński, 2007). Its findings have in turn been verified and corrected by the linguists who were instructed to adjust the mention borders, detected heads and morphosyntactic descriptions.
The data set consisted of 15 texts, about 20 sentences each. All texts contained 5722 tokens, 1644 mentions and 1256 mention chains (including singletons). The major difference from the test set used in the previous, rule-based attempt is the exclusion of zero anaphora: all artificially added zero anaphora tokens were removed. To the needs of the statistical experiment, the training and evaluation data have been encoded in MMAX (Müller and Strube, 2006) format[2] and featured 3 layers: the segmentation layer, the markable layer and the coreference layer.

[2]MMAX2-based environment is currently used in annotation process, see Fig. 4.

```
<markable id="markable_15"
 span="word_60..word_62"
 mmax_level="markable"
 mention_head="świecie"
 sentenceid="1"
 gender="m3"
 number="sg"
 head_pos="subst"
 head_lemma="Świat"
 head_orth="świecie"/>
```

Figure 2: Markable layer: sample mention description

```
<markable id="markable_7"
 span="word_60..word_62"
 mmax_level="coref"
 min_words="świecie"
 coref_set="set_22"
 min_ids="word_62"/>
```

Figure 3: Coreference layer: sample mention description

Data Format The segmentation layer provides information on text tokenization and constitutes the reference to the subsequent layers. The markable layer stores information about mentions – their boundaries (i.e. tokens being part of each mention), sentence number, grammatical gender and number of the mention and detailed information about the mention head (its part of speech, lemma and orthographic form; see example mention representation in Fig. 2).
Since the markable layer lacks evidence about how entities in text are grouped in coreference chains, this information is stored in coreference layer (see Fig. 3).

Evaluation Results For evaluation, the leave-one-out cross-validation method was used. Table 1 presents the results of BART-based coreference resolution compared with our previous rule-based unsupervised system running on the same data. To facilitate comparisons using other mea-

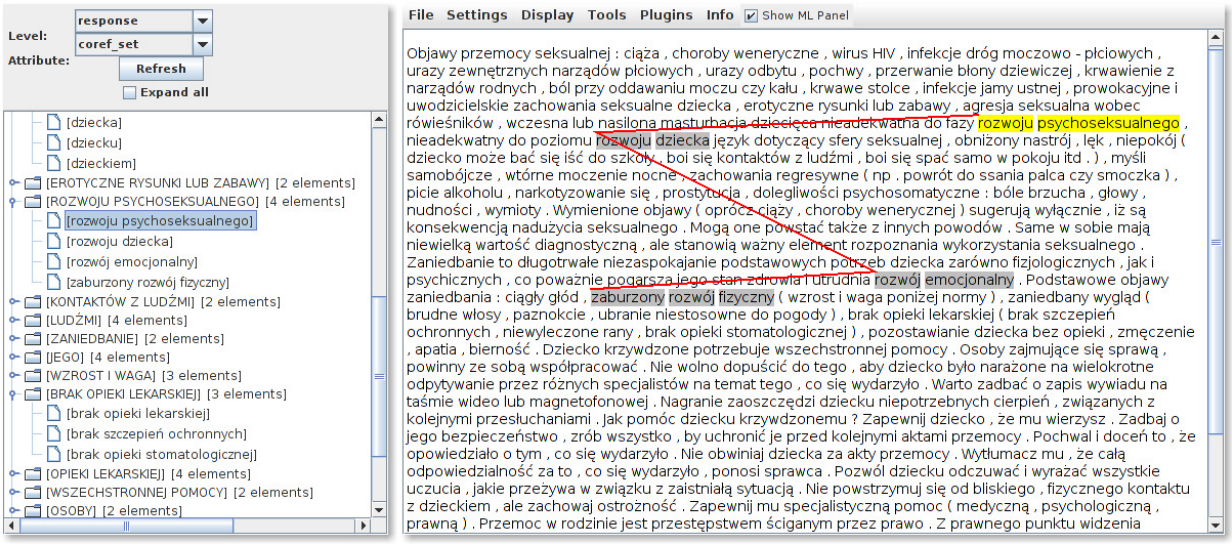

Figure 4: Automatic clustering of coreferent mentions viewed in MMAX2

sures than the only currently (as of October 2011) supported MUC metric, a converter to the SemEval (Recasens et al., 2010b) format has been implemented and the SemEval scorer was used to calculate the remaining values.

MUC (Vilain et al., November 1995) is the metric developed for the Sixth Message Understanding Conference, B^3 (Bagga and Baldwin, 1998) is the B-CUBED metric. CEAFM and CEAFE are both variations of the CEAF (Luo, 2005) metric. CEAFM stands for the mention-based version of it, while CEAFE is the entity-based type. Last metric – BLANC – is the BiLateral Assessment of Noun-Phrase Coreference (Recasens and Hovy, 2010).

The rule-based system achieves higher F1 scores regarding all the measures, but BART comes very close, even having sometimes higher precision, as for BLANC and B^3 measures.

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 1: Comparison of two systems

4. Conclusions and Further Work

The next obvious direction and prerequisite for further experiments is preparation of larger amounts of data since the current size of the evaluation corpus does not allow to capture all features of complex phenomenon of coreference relations. Due to the externally funded CORE project, the annotation of large corpus of Polish coreference is under way.

The current size of the annotated part of the corpus is over 20000 running words, in which already more than 8000 mentions were found. Not only the size, but also the quality of data is going to be better than in the previous experiments, because the annotation guidelines are more precise than at the beginning of the annotation. The diversity of text types is also better in the corpus under annotation, because it maintains their proportion as in the balanced part of the National Corpus of Polish.

However, what needs to be stressed, the sheer large number of training examples would not necessarily improve the score of the system. As results achieved by BART are slightly lower than results of a simple rule-based coreference resolution tool, further language-specific tuning is required, as out-of-the-box solution is not satisfactory. The simplest way would be to change English-specific feature extractors into more generic ones, if possible.

4.1. Preprocessing integration

Another useful task would be to incorporate existing preprocessing tools for Polish into BART, as it would allow to improve the usability of the toolkit and also should increase the language-agnosticism of its core modules. Presented conclusions should encourage researchers to implement coreference resolution pipelines for more languages,

as it is much simpler to build on an existing system than to develop a standalone solution.

4.2. Evaluation *in vivo*

There is no agreement in the community about the best metric for measuring the performance of the coreference resolvers. Because of that, most systems provide their results in terms of multiple metrics. This problem exists because of *in vitro* (intrinsic) evaluation of coreference resolution and it can be tackled by finding an application of such systems as an inner module to solve a different task, which has better established performance metrics.

In context of the international co-operation, the creation of the statistical coreference resolver for Polish, which is the main goal of described work, is intended to create synergy with ATLAS project[3] where an anaphora resolution module is planned to be integrated in the summarization component.

The change of quality of the summaries produced automatically with the different coreference resolution tools would provide a meaningful comparison of them. The tools which are going to be used for that purpose include the adapted version of BART and our simple rule-based system, both described in this paper, but also RARE (Cristea et al., 2002) and Reconcile (Stoyanov et al., 2010) tools, adapted for Polish. The question still exists, how to meaningfully evaluate the quality of automatically created text summaries.

5. References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy approach to Natural Language Processing. *Computational Linguistics*, 22:39–71.

Dan Cristea, Oana diana Postolache, Gabriela-Eugenia Dima, and Cătălina Barbu. 2002. AR-Engine – a framework for unrestricted coreference resolution. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002*, page 2000, Las Palmas, Canary Islands, Spain. Benjamin Publishing Books.

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. pages 25–32.

Ruslan Mitkov, Lamia Belguith, and Małgorzata Styś. 1998. Multilingual Robust Anaphora Resolution. In *Proceedings of the Third International Conference on Empirical Methods in Natural Language Processing (EMNLP 1998)*, pages 7–16, Granada, Spain.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011a. End-to-end coreference resolution baseline system for Polish. In *Proceedings of the 5th Language & Technology Conference (LTC 2011)*, Poznań, Poland.

Maciej Ogrodniczuk and Mateusz Kopeć. 2011b. Rule-based coreference resolution module for Polish. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011)*, pages 191–200, Faro, Portugal.

Adam Przepiórkowski and Aleksander Buczyński. 2007. Spejd: Shallow parsing and disambiguation engine. In *Proceedings of the 3rd Language & Technology Conference*, Poznań.

Adam Przepiórkowski, Rafał L. Górski, Barbara Lewandowska-Tomaszczyk, and Marek Łaziński. 2008. Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco. ELRA.

Marta Recasens and Eduard Hovy. 2010. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, pages 1–26.

Marta Recasens, Eduard Hovy, and M. Antonia Marti. 2010a. A Typology of Near-Identity Relations for Coreference (NIDENT). In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marta Recasens, Lluís Marquez, Mariona Taulé, M A Martí, Véronique Hoste, Massimo Poesio, and Yannick Versley, 2010b. *SemEval-2010 Task 1: Coreference Resolution in Multiple Languages*, pages 70–75. Association for Computational Linguistics.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 156–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Association for Computational Linguistics (ACL) Demo Session*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. November 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pages 45–52.

Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. WEKA: Practical Machine Learning Tools and Techniques with Java Implementations.

[3] Applied Technology for Language-Aided CMS co-funded by the European Commission under the Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 250467).