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.