Coreference resolution in BART

Abstract
In this essay I explore BART, a toolkit for coreference resolution. I test it in a number of different ways, first using its web demo function, and then its experiment function. With the latter, I test the sample data set (provided with BART) in a number of different ways, followed by attempts to test a bigger, Polish corpus. This however turns out to be more difficult than expected, and I leave the first version of this essay without succeeding to do so. Lastly I discuss the results, and lack of results.

1 Introduction
Reference resolution is “the task of determining what entities are referred to by which linguistic expressions”.\textsuperscript{1} Several expressions referring to the same entity \textit{corefer} to the entity, and a reference to an already introduced entity is called \textit{anaphora}. A set of expressions referring to the same entity is called a \textit{coreference chain}, and it is the task of a coreference resolution system to find these coreference chains. With that in mind, I wanted to explore a tool dealing with coreference resolution, and found the open-source modular system BART (\textit{Beautiful Anaphora Resolution Toolkit}).\textsuperscript{2}

BART is designed primarily for English, but it can be used for other languages as well, although with some drawbacks. For example, the preprocessing can only be done with the english presets, so if you are using non-english data, you will need to preprocess the text in some other way. And even then the toolkit performs better on English data, than for data of some other language, due to the its specialization in English.

The system architecture of BART\textsuperscript{3} first takes the text through preprocessing, which consists of, among others, chunking and parsing. After that follows creation of mention objects, based on the preprocessing results. Lastly, a decoder uses one of several available classifiers to learn, and to generate the output: the coreference chains.

BART has two different modes, with a number of different ways of using each

\textsuperscript{1}Jurafsky and Martin (2009, p. 729)
\textsuperscript{2}Also known as Baltimore Anaphora Resolution Toolkit. See \url{http://bart-anaphora.org/}.
\textsuperscript{3}Versley et al. (2008)
one. The most immediate mode is its web demo, where you simply input text, and get chunks, ename, markables, and coreferences. The next function is also a part of the web demo, but with the option of using .txt-files instead of text input, generating the output as XML-documents. This is more suitable if the text is of a larger size. The second and last mode, and the one I have concentrated on, is the experiment function, that runs basic training and testing on a data set of your choice.

2 The web demo

My initial attempts were to simply use the web demo (with text input). These attempts were successful, so I quickly moved over to the other part of the web demo, using a .txt-file with the GNU License Agreement. However, I could not get this function to work. I tried with all the specified commands, I tried with moving the file to another folder and I tried making the file smaller, but nothing worked. In the end I simply tried using the text input function of the web demo, and copied the entire license agreement there.

When it comes to the web demo’s results, I’m not too impressed. I tested this with part of a newspaper article I found online. It got most of the pronoun resolutions correct, but not all of them. For example, in the sentence

   *As in many other Christian countries, many Swedes keep candlesticks with four candles in their homes during Advent.*

the demo paired together *many other Christian countries* with *their*, when the antecedent for *their* should have been *many Swedes*. And in another case, the sentence

   *The star symbolizes the Star of Bethlehem.*

contains references to 2 different stars, whereas the demo linked together *The star* and *the Star*.

3 The experiment function

My first attempts with the experiment mode was to use it with the provided sample corpus. This all went pretty effortlessly, following the instructions of the readme. I also tried changing the classifier, from the default Weka J48 classifier, to a maximum entropy classifier. This gave a slight increase in accuracy, but also doubled the time it took to classify. Following this, I also tried testing the same data set with the generic language settings, which, not surprisingly, gave a slight decrease in accuracy. The results of all the tests of the sample data set can be seen at table 1.

---

4 The Local (2013)

5 7 506 tokens from the Wall Street Journal.
I also wanted to try using BART with another, bigger corpus. This was a task that proved to be more difficult than I had believed. I could find a corpus\(^6\) in the correct format (\texttt{MMAX}), and I could seemingly successfully preprocess it. But any coreferences could not be found during the preprocessing or testing, and thus the testing returned invalid results. The reason for this is most probably that BART does not have any support for the Polish language. I will not explore the topic of the Polish compatibility of BART further, and I refer to Kopeć and Ogrodniczuk (2012) for more information on the subject.

I can not fully compare my results with the results from other studies, since

---

\(^6\)The Polish Coreference Corpus, see
http://zil.ipipan.waw.pl/PolishCoreferenceCorpus
different testing material has been used. I can however observe general results. Looking at the results for the sample data set, the F1-scores get to approximately 60% in all cases. This is close to the scores reported in Versley et al. (2008). Results of other systems in Versley et al. (2008) and Charniak and Elsner (2009) are also around that number.

4 Discussion

Charniak and Elsner (2009) tried to include BART in their evaluation of their own system compared to a few other systems, but were unable to do so because they could not get the program to work. But being that the year of the publication is 2009, and the year of the current version of BART is 2013, a lot might have happened since then. And the program is actually working now, but I feel that it needs some improvements when it comes to usability.

One of the purposes of BART was to “significantly lower the entrance barrier for researchers interested in coreference resolution”.7 I believe that the authors with this were referring to the web demo rather than the experiment function. While the web demo, apart from the non-working .txt-file input, might have been helpful, I did not feel that the experiment function was. The usage of the experiment function was alike other, similar systems, but I felt that the MMAX format made using BART with a new corpus difficult.

The fault for the MMAX format is not entirely BART’s, though. The MMAX2 Annotation Tool8 was the only available MMAX annotation tool I could find, and this too did suffer from bad usability, when I tried to prepare the files of the Polish corpus in it. But for the final version of this essay I intend to dig deeper into preprocessing the Polish corpus, and hopefully I will be able to successfully test it in BART. After some initial tries I have my doubts, though.

---

7 Versley et al. (2008)
8 Müller and Strube (2006)
References


