PFL054 2008/09

Term project

Coreference Resolution

The aim of the project is to solve „the best you can“ (with regards to the suitable measures) binary classification task of the resolution coreferential and noncoreferential word pairs.

What is the Coreference?

Example: ***Father*** *claimed disliking opera-going.* ***He*** *was mainly angry with the style of opera singing.*

The words *father* and *he* denote the same individual – a father of the major hero from the selected book. Both words take part in the process called **reference**, where you can use more different expressions (words, phrases) to refer to individuals, subjects or situations of the real world. A natural language expression used to perform reference is called a **referring expression**, and the entity that is referred to is called the **referent**. Thus, *father* and *he* are referring expressions, and the real-word man is their referent. Two referring expressions that are used to refer to the same entity are said to corefer, the relationship between them is **coreference** and these two expressions form a **coreferential pair.**. Usually the second expression (mentioned later in the text) in a coreference pair (*he*) is called **anaphora**, the first (mentioned earlier in the text) one (*father*) is **antecedent**.

Of great importance to the study of coreference is **annotation** of coreferential pairs in the text according to the selected methodology. We use the methodology designed in the Prague Dependency Treebank (PDT, http://ufal.mff.cuni.cz/pdt2.0/index.html). See the sample at http://ufal.mff.cuni.cz/~hladka/sample.html.

The knowledge of coreferential pairs is an important part of many applications in the natural language processing – e. g. information retrieval, document summarization, machine translation. We work with the enormous volume of texts so it is practically impossible for human beings to go through them manually. That is why an automatic procedure which is capable to detect coreferential pairs in the text is needed – we speak about an **automatic coreference resolution**. For Czech language, there have already been done the experiments on the automatic coreference resolution within the PDT methodology. For more details, see:

1. Nguy Giang Linh: Návrh souboru pravidel pro analýzu anafor v českém jazyce (A set of rules for anaphora resolution in Czech), MFF UK 2006. (http://ufal.mff.cuni.cz/~hladka/ML/aca-diplomka.pdf)
2. Nguy Giang Linh; Žabokrtský, Z.: Rule-based approach to pronominal anaphora resolution applied on the Prague Dependency Treebank 2.0 data. In Proceedings of DAARC 2007 (6th Discourse Anaphora and Anaphor Resolution Colloquium). http://ufal.mff.cuni.cz/~zabokrtsky/papers/daarc-2007.pdf

Data format for training & testing

Data are divided into the train and evaluation (test) sets: `train.csv` and `etest.csv`. Each row in data files corresponds to one word pair. In case of **anaphora** and **candidate-antecedent**, the pair is classified as being coreferential, i.e. a positive instance. The data are prepared to apply machine learning methods so there are noncoreferential (i.e. negative) instances as well (i.e. a pair of **anaphora** and **candidate-non-antecedent)**. There is at least one noncoreferential pair for each coreferential pair (and vice versa) in the data. There are around 14% coreferential pairs out of all pairs in the data.

There are 55 comma-separated values on each row. The first 54 values are features which you can use in the machine learning methods. Values and names of the features are described in `all.names` and `anaphora.names`. The very first feature is a technical one and serves as an indicator of the anaphora. The rest of the features are divided into the categorical and continuous type. All possible values of the categorical features are listed in `all.names` (comment: data for your experiments presents just a small portion of the complete data, so some of the values need not to occur at all). Obviously, there is no such list for continuous features. For more details about the features read Czech or English version of the reference manual ((http://ufal.mff.cuni.cz/pdt2.0update/). Last value in a row (just before fullstop, [0|1]) codes an information about the noncoreferential or coreferential relation of a given pair.

Example: Two instances: noncoreferential and coreferential for the same anaphora:

tundefcmpr9410undef001undefp10s2a0,fem,sg,fem,sg,1,1,0,0,PAT,v,CNCS,v,0,1,0,PAT,ACT,0,Obj,empty,0,N,empty,N,empty,F,empty,S,empty,4,empty,undef,empty,undef,empty,undef,empty,1,1,1,0,1,0,1,1,11,3,f,t,0,0,0,0,**0**.

tundefcmpr9410undef001undefp10s2a0,fem,sg,fem,sg,1,1,0,0,RSTR,v,CNCS,v,0,1,0,ACT,

ACT,1,Sb,empty,0,P,empty,4,empty,F,empty,S,empty,1,empty,undef,empty,undef,empty,undef,empty,1,1,1,1,1,1,1,1,15,4,t,t,1,0,0,0,**1**.

The goal in details

The aim is to detect in the best way (in a sense of suitable measures, see below) both coreferential and noncoreferential pairs, i.e. a binary classification task. You must apply AT LEAST TWO METHODS presented during the lectures and the seminars. Setting up (redefining and setting the number of) the features used for classification is a part of the project.

Methods should be trained ONLY on training data available. Reliable results have to consist of information on the error rates expressed by the suitable measures (accuracy, confusion matrix, precision, recall, F-measure) on the testing data. Comparison of the results on training and testing data is welcome.

Deadlines

December 13, 2008, 24:00

You will need to turn in electronically to {hladka, schlesinger}@ufal.mff.cuni.cz:

- a **programming** (R code)
- a **short report** describing methods, results and comments. A short report should be 1 page (A4) in length, excluding figures.

December 15, 2008 , 10:40

- 10 minutes presentation during the seminar

February 20, 2009, 24:00

You will need to turn in electronically to {hladka, schlesinger}@ufal.mff.cuni.cz:

- a **final programming** (R code)
- a **final report** written according to the guidelines specified at the http://ufal.mff.cuni.cz/~hladka/ML.html -> Project
- a proposal of date and time of the meeting with Pavel Schlesinger

!!! YOUR TASK MUST BE DONE BEFORE FEBRUARY 20, 2009. AFTER THAT YOU WILL NOT GET “A SIGNATURE”. YOU CAN TAKE THE EXAM BEFORE YOU FINISH THE FINAL PROJECT. HOWEVER, TO GET A FINAL GRADE, YOU NEED TO FINISH THE FINAL PROJECT (I.E. TO HAVE “A SIGNATURE”)!!! Pavel Schlesinger will consult with you on the problems you will meet while doing the project. E-mail him to make appointment.