PFL054 2009/10

Term project

Verb-sense disambiguation

The aim of the project is to address a task of choosing the proper sense of selected verbs in their given context, i.e. verb sense disambiguation task.

Description

Verbs are central elements of clauses with strong influence on the whole sentence. Therefore the semantic analysis of verbs plays a key role in the analysis of natural language.

Disambiguation of verb sense is of great importance for applications of natural language processing such as machine translation. A highly ambiguous Czech verb *dát* is really illustrative example:

CZ: Petr **dal** Janě knihu.	EN: Peter **gave** Jane a book.
CZ: Petr **si dal** klíče do kapsy.	EN: Peter **put** his keys in his pocket.
CZ: Petr **si dal** Guiness do půllitru.	EN: Peter **ordered** a pint of Guiness.

Data

The instances correspond to the occurrences of the verb in the data (#occurrences = #instances) and each instance is described by the features. Features present a key word of any machine learning algorithm. In our task, features describe different information about the context of the verb within the sentence. We distinguish the features of five types – morphological (m, n), syntax-based (s), idiomatic (i), information about the animacy (a), information from the WordNet (w). So for the verb (**verb**) and for the feature types ([**mnsiaw**]), twelve data files are generated from the particular language resources: **verb.[mnsiaw].names** with description of the features and **verb.[mnsiaw].data** with the features from the data. Additionally, two more data files are available: **verb.x.names** with a list of all possible senses of the verb and **verb.x.data** with the disambiguated senses of the verb (i.e. instances) in the data.

The goal in details

The aim is to choose the proper sense of the SELECTED Czech verbs. You are supposed to address this task in two steps:

1, By December presentation (see below Deadlines), you will disambiguate the proper sense of Czech verb *přihlížet* that has *****two***** possible senses according to the Czech valency lexicons (for translation see Seznam Czech-English dictionary at http://slovnik.seznam.cz/?q=p%C5%99ihl%C3%AD%C5%BEet&lang=en_cz)

You will be provided with the training data files **train.[mnsiawx].data**, **train.[mnsiawx].names** and the test data file **test.[mnsiawx].data** (generated from the files přihlížet. [mnsiawx].[data|names]).

You must apply either the decision trees algorithm or the Naïve Bayes algorithm. But you are not allowed to make your own choice ;-(We will send you a message which algorithm you will apply.

The results have to be presented by the suitable measures (accuracy, confusion matrix) on the test data. Cross-validation and bootstrapping are welcome.

2, By February 19 (see below Deadlines), you will disambiguate the proper sense of two verbs – *přihlížet* and *odpovídat*. You have already had the data for the verb *přihlížet*. Data (of the same format as in 1,) for the verb *odpovídat* of *****three***** possible senses are posted at http://ufal.mff.cuni.cz/~hladka/ML/EXAM_2010/data/data2_project.zip. You must apply at least three machine learning algorithms you met during the lecture. In this case, the choice is up to you;-) Setting up (redefining and setting the number of) the features used for classification is a part of the project. Do not forget to compare results of the different methods.

Deadlines

December 6, 2009, 24:00

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

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

December 7, 2009 , 15:40

- 10 minutes presentation during the practice

February 19, 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 posted at the http://ufal.mff.cuni.cz/~hladka/ML.html -> Project, suggestion for date and time when you will meet Pavel Schlesinger to consult your solution. Please take into account that your final report is very likely to be consulted with PS and changed afterwards. (Given the experience from the previous years, approx. 2 updates are expected.)

!!! YOUR FINAL REPORT MUST BE SUBMITTED BEFORE FEBRUARY 19, 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”)!!!