Introduction to Machine Learning

Default Term Project Specification

NPFL054 2014/15

Native Language Identification

The aim of the default term project is to solve „the best you can“ a classification task. Your classifier will predict the first language (L1) of English essay's authors with the help of The ETS Corpus of Non-Native Written English (former TOEFL11) where L1, prompts (topics) and proficiency levels are provided. This term project is inspired by the NLI-2013 Shared Task.[1]

1. Project Assignment

Task

Having a collection of English essays written by non-native speakers, the goal is to predict a native language of the essays' authors. Languages L1 are known in advance. We have a collection of English essays for which L1 is known – The ETS Corpus of Non-Native Written English, thus we formulate this task as a classification task addressed by using supervised machine learning methods.

Default setting

The students will choose one of the following default settings to address the default term project. The selected setting must be used in their experiments. At the same time, there are no other restrictions. The students are free to use any machine learning concept that they can see useful for increasing the performance achieved with their default setting.

(A)

- **Default features:** word *n*-grams (n=1, 2) and part-of-speech tag n-grams (n=1, 2)
- **Default algorithm:** Support Vector Machines

(B)

- **Default features:** lemma n-grams (n=1,...,2) and part-of-speech tag n-grams (n=1, 2)
- **Default algorithm:** Logistic Regression

[1] http://www.cs.rochester.edu/~tetreaul/naacl-bea8.html#nli

(C)

- **Default features:** character n-grams (n=1,...,5)
- **Default algorithms:** Support Vector Machines and Logistic Regression

Evaluation: 10-fold cross-validation with the sample *Train* ∪ *DevTest*. Report accuracy and confidence intervals.

When you finish all your work and your program codes, you will submit your final solution in the form of a detailed report. The report should contain both the description of the methods used (including the description of parameters tuning) and the analysis of the results. Conclusion of your final report will include your choice of the best model you have developed. Your best classifiers will be evaluated on an evaluation data set that will be hidden from you until you submit your final classifier.

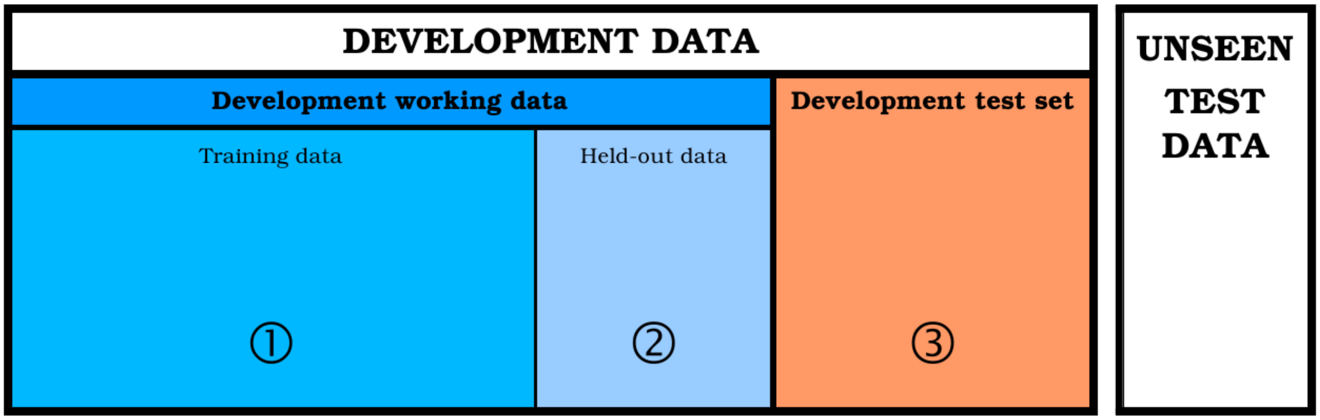
2. Data

A new publicly available corpus of non-native English writing called The ETS Corpus of Non-Native Written English (https://catalog.ldc.upenn.edu/LDC2014T06) consists of essays on eight different topics written by non-native speakers of three proficiency levels (low/medium/high); the essays' authors have 11 different native languages, namely Arabic (ARA), Chinese (CHIN), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR). The corpus contains 1,100 essays per language.

Attention! The ETS corpus is not free and was legally bought by the Institute of Formal and Applied Linguistics, Charles University in Prague. You are approved to use it exclusively for this project and not for any other purpose.

You will be provided with the development data, namely training set *Train* and test set *DevTest,* in the `nli-data.zip` zip file

- `index-training.csv ##` development working data
- `index-dev.csv ##` development test data
- `responses_tokenized/ ##` tokenized essays
- `prompts/ ##` (topics) for the essays
- `README`



When you develop your classifiers, you will use the development test set both to evaluate your classifiers and to tune their parameters. Once you have finished your parameters tuning, you will choose the best model and use all the annotated data (i.e. *Train* U *DevTest*) to train your final, "best" classifier. This final classifier will be submitted and then evaluated on the unseen test set.

3. Milestones

(1) Data delivery & task assignment

You will be assigned with one of the three settings described above on November 27th. In addition, you will be given *Train* and *DevTest*.

(2) Issues to address before short report presentation

- **by Dec 3**
  - Focus on data analysis. For each L1 language,
    - get statistics on the number of tokens and sentences from *Train* U *DevTest*
    - study the most frequent and least frequent tokens in *Train* U *DevTest*
    - get statistics on the distribution of prompts and learners' proficiency levels from *Train* U *DevTest*
    - define features you will experiment with (include “your” default features)
    - do preprocessing, if needed (for ex. if you want to work with part of speech tags)
  - Write the first draft of your short report
    - Describe shortly “your” algorithm(s)
    - Report the statistics
  - Send the draft to hladka@ufal.mff.cuni.cz
- **by Dec 15**
  - Extract default feature vectors
  - For each feature, calculate its information gain on *Train* U *DevTest*
  - Train your very first classifier on *Train* and get its accuracy on *DevTest*
  - Write the final version of your short report:
    - include statistics on your feature vectors
    - describe the classifier, mainly focus on learning parameters, and comment its accuracy

(3) Short report submission

Before the presentation, you will send us your short report by **Monday, December 15th, 2014, 12pm**. You will need to turn in electronically to hladka@ufal.mff.cuni.cz:

- Your short report (.pdf) describing methods, results and comments. The short report should be 1 page (A4) in length, excluding figures and tables.

- Your programming (R codes).
- Your slides (.pdf) prepared for short oral presentation (15 minutes at maximum).

(4) Short report presentation

will take place at the lab session on **Wednesday, December 17th, 2014**.

(5) Final report submission

You do not have to present your final report publicly. Instead, you will defend your work individually. You should be able to explain all details and discuss the choice of your solution in personal conversation with the teacher. The deadline for your final report is Friday, **February 15th, 2015**, 12pm. You will need to turn in electronically to hladka@ufal.mff.cuni.cz:

- Your final report (.pdf)
- Your final programming (R code)

Filename convention

Everytime when you submit your work, please send always just ONE zip file, and follow the filename convention: whole package: "`YourLastName.ml-project.2014-15.zip`"; your R-scripts inside the package: "`YourLastName.ScriptNameofYourChoice.R`"; your report inside the package: "`YourLastName.report.[short|final].pdf`"

For inspiration, see the student reports from the past posted at https://ufal.mff.cuni.cz/course/npfl054/final-projects (Nice Student Reports)

Remember well, that your work MUST be done by **February 15th, 2015**. After that deadline you cannot get “a signature”. You can take the exam before you finish the final project. However, to get a final grade and the credit, you need to finish the project (i.e. to get “a signature”). Before you submit your final report you will have opportunity to consult Barbora Hladka about the problems you meet while working on the project. Do not hesitate to e-mail her to make appointment.