Master thesis proposal

Ion Madrazo Azpiazu

Wednesday 25th November, 2015

1 Introduction

- **Context** What is a readability score?
- **Audience** Text and book recommendation for children or for language learners. Automatic text simplification, for summary generation, for people with reading difficulties. Literacy assessment. Political or medical document complexity assessment.
- **What is the problem?**
  - Historically used metrics are not precise enough. They base they work on very simple or shallow features, such as word length or sentence length.
  - Tools developed for readability prediction are usually monolingual. They tend to be monolingual because the tools and datasets used are so too.
- **What do we propose?** An exploration of new features that fit most of the languages and that fit specific languages. A multilingual tool that is able to detect the input language on the fly and use the best set of features for that specific language for prediction.
- **In doing so we will contribute to**
  - **An application** that will help people with different profiles in selecting texts and books in different languages.

  - **Several datasets**, that will be created as a byproduct of the development and the testing of the applications.

- **Case study** Even if the application will be able to work in many more languages, for practical purposes, the application will be tested in three different languages. English, for state of the art comparison purposes and as reference of germanic languages. Spanish, as a reference for latin languages, and Basque as an example of a non-indoeuropean language.

2 Related work

2.1 Historical readability measures

Description of basic readability scores. When and where were they used? Fleisch etc...

2.2 General State of the art

2.3 State of the art for English

2.4 State of the art for Spanish

2.5 State of the art for Basque

2.6 State of the art for multilingual predictors

3 Methodology

3.1 Pipeline decription

Describe the whole process for prediction. Texts processing, feature extraction, feature processing, prediction.

3.2 Text processing

Description of Freeling NLPTools, which modules are we using. Description of each module with a general description and examples. Generate hypotheses of why this module should add valuable information to the text.

3.3 Feature extraction

Description of all the features used. Why should this feature be valuable, give hypotheses and intuition behind the use of each feature. Give examples when needed.

3.4 Feature processing and selection

Describe algorithms used for feature processing and selection, why should they help get better results?

3.5 Learning and prediction

Describe algorithms for learning and prediction. Pros an cons of each algorithm, why should this algorithm adapt better to our problem?

4 Evaluation

4.1 Datasets

Information about how we get and extract the datasets.

4.1.1 English

- Lexile
- List all for proposal...

4.1.2 Spanish

- Lexile
- List all for proposal...

4.1.3 Basque

- Ikasbil

4.2 Metrics

- Error rate, accuracy
- Adjacent accuracy, double adjacent accuracy...
- Average error distance

4.3 Tests

- Which features add the most value? Correlation, information gain etc.
- Do features correlate similarly with the readability score for each language?
- Feature preprocessing, does it help?
  - Discretization
  - Feature subset selection techniques
- Comparison of learning models, which learning model fits best the problem?
  - KNN
  - Bayesian models
  - SVM
  - Neural network
  - Ordinal classification
- **Comparison** of the system vs **baselines** such as fleish for each language individually.
- Comparison **vs state of the art** systems for each language.
- Multi vs monolingual
- If we take a bilingual corpus, does the system predict same values? And if we take a text and translate it to another language? Does the readability values mantain using an automatic translator?