Master thesis proposal

Ion Madrazo Azpiazu

Wednesday 18th November, 2015

1 Introduction

- What is a readability score?
- What are the uses of a readability score?

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

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

4.1.2 Spanish

- Lexile

4.1.3 Basque

- Ikasbil

4.2 Tests

- **Comparison** of the system vs **baselines** such as fleish for each language individually.
- Comparison **vs state of the art** systems for each language.
- 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
- 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?