Master thesis proposal

Ion Madrazo Azpiazu

Thursday 14th January, 2016

1 Introduction

Reading is an important skill of not only in the educational context, but in daily tasks. People that get good understanding from the texts they read, tend to take better decisions when they face medical or educational problems, which can lead to better career opportunities in the future. Studies [1] show that even medical resources that are supposed to be suited for average readers, tend to have too specialized vocabulary tha even well educated adults have trouble understanding. Therefore, it is important no only to provide students with encouraging texts for improving their reading skills, but also to be able to produce texts, in medical or legal contexts, that are simple enough to be understood by people with average or poor reading skills . That is where the readability assessment formulas take place, providing a simple way of assessing a text's complexity level, so that they can be adequate to a public as broad as possible.

A readability score refers to the degree of ease with which a reader can understand a given text, which is usually conditioned by features such as vocabulary or syntax. Historically, teachers have been the main stakeholders of the use of readability formulas, making use of it for obtaining new materials for their courses and curriculum design. However, lately, readability scores have been discovered to have more uses than the ones in academic environments. Automatic text simplification or summarizing for people with reading difficulties, literacy assessment, or even political and medical document complexity assessment are examples of the benefits a readability score can provide.

However, historically used readability measures have shown poor performance. Due to the lack of computational power readability measured used to be based on shallow features such as average sentence or word length. Even if those features have showed to be a good baseline for readability assessment, they are not precise enough for certain tasks. Recent developments in the area have made use of both machine learning and natural language techniques, that make good use of the current computational capabilities, showing an significant improvement in precision. However, these tools tend to make use of language specific features that make them only useful for one specific language.

The system we propose would try to solve the two issues discussed above. We aim to develop a multilingual tool that is able to detect the input language of a text on the fly and use the best set of features for that specific language for predicting its readability. This tool would both significantly improve precision over the baseline and be multilingual.

In doing so we will contribute to (1) the development of **an application** that will help people with different profiles selecting texts and books in different languages, (2) an **analysis** of current features and tools used in the literature and (3) several **datasets** that will be created as a byproduct of the development and the testing of the application.

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 Thesis statement

- Develop a multilingual readability predictor taking advantage of machine learning techniques and features extracted using natural language processing techniques.
- Develop a survey of currently used readability tools and methods, together with a comparison of features and their importance in the read-

ability prediction for each language.

3 Related work

3.1 Historical readability measures

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

3.2 General State of the art

3.3 State of the art for English

3.4 State of the art for Spanish

3.5 State of the art for Basque

3.6 State of the art for multilingual predictors

4 Methodology

The proposed method relies in two different areas of data science, Natural language processing and machine learning. Advantage of one or both areas is taken in each of the steps that conform the pipeline of the algorithm explained below.

4.1 Pipeline description

The pipeline of the algorithm if composed by the following steps: Texts processing, feature extraction, feature processing and prediction. A visual description of the general pipeline of the system can be seen on figure 1. A more in-depth explanation of each step can be seen in the following sections.

4.2 Text processing

The text processing step is the step where the raw text is given structure and, therefore, value. This structure and information will later be used for

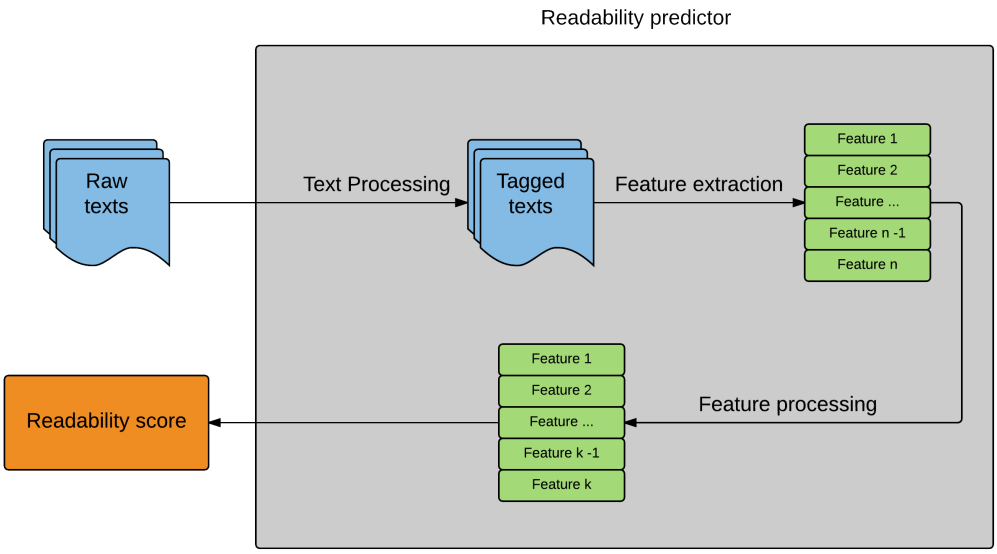

Figure 1: General pipeline

extraction features that will help the system predict a readability score.

The tool that has been chosen for natural language processing is Freeling NLP [**?**]. Freeling is an open source Natural language processing library that supports 11 different languages. The tool solves common NLP tasks, such as, Tokenization, sentence detection, Part of speech tagging or dependency parsing. Each of this processes will be helpful for building certain features later.

The **tokenization** is the base module for any NLP processing. Tokenization refers to taking a raw text and normalizing it into pieces that make text processing possible. This will also make possible, to implement tradition shallow features such as, FleschKincaid [**?**].

The **Part of speech** analysis determines the function each token has in the sentence. This, together with **dependency parsing** techniques, make possible the analysis of syntactic structures in the sentences.

Other tools outside Freeling, such as **WordNet** or **Latent semantic analysis** techniques, will make possible to analyses texts at semantic level, for detecting structures that refer to concepts rather than to tokens themselves.

4.3 Feature extraction

This section describes the features proposed for the system. These features range from the most simple and commonly used ones such as the shallow features, to a more complex set of features such as the ones base on semantics.

Shallow features

Part of Speech tags

N-grams

...

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.

4.4 Feature processing and selection

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

4.5 Learning and prediction

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

5 Evaluation

5.1 Datasets

Information about how we get and extract the datasets.

5.1.1 English

- Lexile
- List all for proposal...

5.1.2 Spanish

- Lexile
- List all for proposal...

5.1.3 Basque

- Ikasbil

5.2 Metrics

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

5.3 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
  - Regression (Adding a sense of order in class values)
  - Ordinal classification (Adding a **stronger** sense of order in class values)

- **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 maintain using an automatic translator?

References

[1] Ahmed MS Ibrahim, Christina R Vargas, Pieter GL Koolen, Danielle J Chuang, Samuel J Lin, and Bernard T Lee. Readability of online patient resources for melanoma. *Melanoma research*, 26(1):58–65, 2016.