Master thesis proposal

Ion Madrazo Azpiazu

Tuesday 19th January, 2016

1 Introduction

Reading is an important skill in the academic environment, a skill that can be deciding for a student's educational opportunities and their career [?]. Therefore, giving students students encouraging texts to read during each of the stages of their lives is essential. Even outside the educational environment, reading plays a important role. It is primordial for people that get good understanding from the texts they read, specially when they face medical or legal issues. Understanding a legal or medical document properly, can lead the reader to taking a better decision and having more confidence about it. However, studies [Ibrahim et al., 2016] [Petkovic et al., 2015] [Patel et al., 2015] show that even medical resources that are supposed to be suited for average readers, tend to be too specialized and even well educated adults have trouble understanding. Being able to produce simple texts, that can be understood by people with low reading skills seems imperative.

A readability score refers to the degree of ease with which a reader can understand a given text and score which is usually determined by a readability formula. Historically, teachers have been the main stakeholders of the use of readability formulas, making use of them 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 [Štajner et al., 2015] [Saggion et al., 2015] or summarizing for people with reading difficulties [Feng, 2009a], book recommendation [Pera and Ng, 2014], literacy assessment [Weiss et al., 2005], or even legal [Ogloff and Otto, 1991] and medical document complexity assessment [Ibrahim et al., 2016] [Petkovic et al., 2015] [Patel et al., 2015] are only a few examples of the benefits a readability score can provide.

During a few decades, traditional formulas such as Flesh [Flesch, 1948], Dale-Chall [Chall and Dale, 1995] and Gunning FOG [Albright et al., 1996] became very popular among the educators for manually determining text difficulty. Most of those formulas made use of *shallow features* such as, average word length or average sentence length, which provided a simple way of determining a texts complexity, despite lacking precision in some cases. Cases such as the one claimed by (David and kantor 1982)(manually cited) where nonsense text could be classified as simple to read, just because it contained short and frequently used words. Cases like that encouraged researchers to study better methods of prediction. Therefore, over time, the traditional shallow formulas started to leave place to more complex formulas [François and Fairon, 2012] [Aluisio et al., 2010], that began to put together both natural language processing and machine learning techniques. The new formulas usually continued using the shallow features mentioned above, but added new more complex features based on syntax or semantics of the text. With the addition of new features, the tools became more precise, but more constrained regarding the language. They started to use more and more language dependent techniques and tools, which made the systems difficult to adapt to other languages than the one they were designed for, making the multilingualism that was possible in the early stages disappear.

Having seen the lack of multilingualism among the state of the art systems for readability prediction, we propose to develop a multilingual readability assessment tool . This tool should both show results comparable to monolingual state of the art systems, and maintain the multilingualism the early tools in the readability field had. For doing so we will make a exploration of the features and methods used in literature, and adapt them to be multilingual. Furthermore, we will develop novel features and that will aim at being useful for each of the languages, individually or collectively. Finally, we will analyze the effect each of the tested features have regarding readability, determining by this, what features make a text readable generally and specifically for each language. In doing so, we will produce a system that will adapt itself to the input text language, and use an adequate subset of features for that certain language for giving a prediction, creating, to the best of our knowledge, the first multilingual readability assessment system.

It is important to note, that for practical purposes, the application will only 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

Let's discuss about this section next meeting, looks like a big repetition of paragraph 4 to me.

We aim to develop a multilingual readability predictor taking advantage of machine learning techniques and features extracted using natural language processing techniques. As a secondary goal, we will survey the features and methods currently used in the state of the art, and create a comparison features and their importance in the readability prediction for each language. As a byproduct of the development and testing, we will create various datasets that can be of good use for other researchers in the area.

3 Related work

In the recent years, different Readability Assessment (RA) systems have been developed with high diversity regarding both languages and features.

I have spent some time looking for multilingual tools and it looks that there is nothing. Where should I talk about that? Maybe in previous paragraph? Or right at the end of the RW?

For **English**, [Feng et al., 2010] presented a comparison of the common readability features used for English. [Aluisio et al., 2010] aimed their system for evaluating text simplification methods with a system, that made use of some more elaborated features such as ambiguity in terms. [Feng, 2009b] oriented their system for assessing the difficulty level of a text for people with intellectual disabilities, developing some features that were intended to detect how well a text was structured. Will find a couple more.

For **Spanish** several systems [Štajner and Saggion, 2013] [Drndarević et al., 2013] have focused their work on text simplification and its evaluation using classical readability formulas. Formulas such as, SSR [Spaulding, 1956] based on sentence length and number of rare words per sentences or LC and SCI [Anula, 2007] based on density of low frequency words in text. Let's try to find another one. Not so easy here.

For **Basque**, for the best of our knowledge, only one system have been developed. Due to the fact that Basque is considered a minority language and shares very little similarities with the most spoken languages. Very little research have been done in the area. Therefore, currently, Errexail [Gonzalez-Dios et al., 2014] is the only system created for Basque readability assessment. This system was aimed for text simplification purposes and was developed to predict two different values, simple or complex. The aim for this was to detect which texts needed some simplification and which texts did not. The system makes use of simple features mostly based on ratios of common Natural language processing tags.

For **Chinese**, [Chen et al., 2011] developed a RA system only based on lexical metrics based on the TF-Idf measure. This metric in conjunction with a mutual information measure was able to determine which terms were most relevant for each of the readability levels. These terms were afterwards used to predict the level of readability for the inputs texts. However, this technique was not topic independent, as once trained for a certain topic the terms were no longer useful for other topics. Previously, [Collins-Thompson and Callan, 2004] developed a system that already tried to solve, the topic dependence problem for Chinese. This system was based on Tf-Idf too and as the authors stated, removing some top scoring words of the Tf-Idf ranking, lead the system to be more independent of the topic. Apparently, the top scoring words were highly specialized words to the topic selected for training.

For **Arabic**, [Al-Ajlan et al., 2008] developed a readability assessment told based on only two features. The features were based on simple ratios based on sentence,terms and letter counts. Those, features were used with a SVM classifier in order to be able to classify text as simple or complex.

For **Italian**, [Dell'Orletta et al., 2011] presented a readability assessment system aimed for text simplification. Since the text simplification tool the authors were developing was based on sentences. The authors of this system decided,that rather than developing a system for determining text readability, their system would work at sentence level. Therefore, the text simplification tool, would have more information of which sentence needed simplification and which did not. The model generated for sentence level is shown to be generalizable to full text level, by the use of simple averages. The more complex sentences a text have, the more probabilities it have to be complex in overall.

For **French**, [François and Fairon, 2012] developed a readability assessment system with the foreign language learners in mind. The objective was to determine which features were more important for a foreign language learner to understand a text. In addition, they provided a metric new to the area called adjacent accuracy that tried to measure systems' performance in a more accurate and relevant.

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

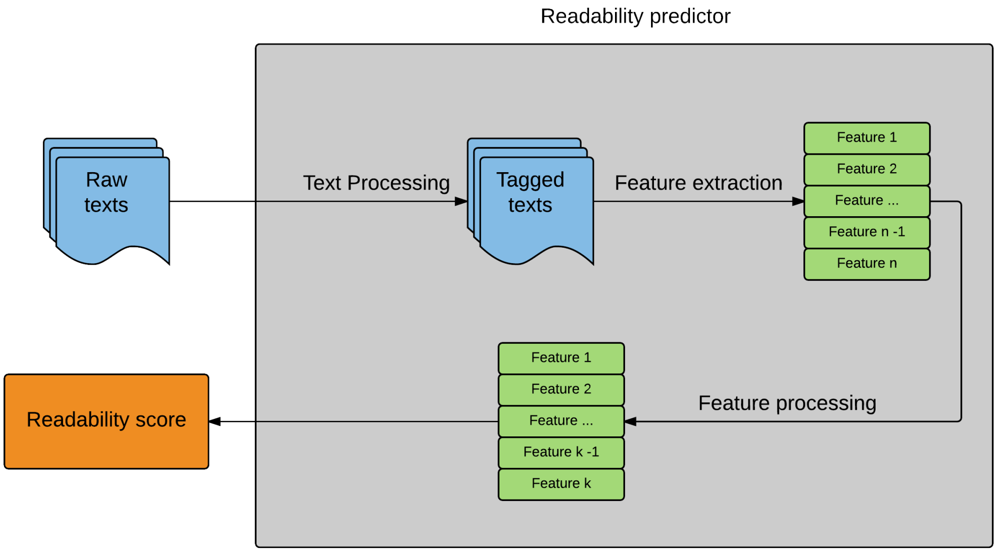

Figure 1: General pipeline

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

[Al-Ajlan et al., 2008] Al-Ajlan, A. A., Al-Khalifa, H. S., and Al-Salman, A. (2008). Towards the development of an automatic readability measurements for arabic language. In *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, pages 506–511. IEEE.

[Albright et al., 1996] Albright, J., de Guzman, C., Acebo, P., Paiva, D., Faulkner, M., and Swanson, J. (1996). Readability of patient education materials: implications for clinical practice. *Applied Nursing Research*, 9(3):139–143.

[Aluisio et al., 2010] Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics.

[Anula, 2007] Anula, A. (2007). Tipos de textos, complejidad lingüística y facilicitación lectora. In *Actas del Sexto Congreso de Hispanistas de Asia*, pages 45–61.

[Chall and Dale, 1995] Chall, J. S. and Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

[Chen et al., 2011] Chen, Y.-H., Tsai, Y.-H., and Chen, Y.-T. (2011). Chinese readability assessment using tf-idf and svm. In *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, volume 2, pages 705–710. IEEE.

[Collins-Thompson and Callan, 2004] Collins-Thompson, K. and Callan, J. P. (2004). A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200.

[Dell'Orletta et al., 2011] Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83. Association for Computational Linguistics.

[Drndarević et al., 2013] Drndarević, B., Štajner, S., Bott, S., Bautista, S., and Saggion, H. (2013). Automatic text simplification in spanish: a comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer.

[Feng, 2009a] Feng, L. (2009a). Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS Accessibility and Computing*, (93):84–91.

[Feng, 2009b] Feng, L. (2009b). Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS Accessibility and Computing*, (93):84–91.

[Feng et al., 2010] Feng, L., Jansche, M., Huenerfauth, M., and Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 276–284. Association for Computational Linguistics.

[Flesch, 1948] Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221.

[François and Fairon, 2012] François, T. and Fairon, C. (2012). An ai readability formula for french as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477. Association for Computational Linguistics.

[Gonzalez-Dios et al., 2014] Gonzalez-Dios, I., Aranzabe, M. J., de Ilarraza, A. D., and Salaberri, H. (2014). Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING*, volume 2014.

[Ibrahim et al., 2016] Ibrahim, A. M., Vargas, C. R., Koolen, P. G., Chuang, D. J., Lin, S. J., and Lee, B. T. (2016). Readability of online patient resources for melanoma. *Melanoma research*, 26(1):58–65.

[Ogloff and Otto, 1991] Ogloff, J. R. and Otto, R. K. (1991). Are research participants truly informed? readability of informed consent forms used in research. *Ethics & Behavior*, 1(4):239–252.

[Patel et al., 2015] Patel, C. R., Sanghvi, S., Cherla, D. V., Baredes, S., and Eloy, J. A. (2015). Readability assessment of internet-based patient education materials related to parathyroid surgery. *Annals of Otology, Rhinology & Laryngology*, page 0003489414567938.

[Pera and Ng, 2014] Pera, M. S. and Ng, Y.-K. (2014). Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 9–16. ACM.

[Petkovic et al., 2015] Petkovic, J., Epstein, J., Buchbinder, R., Welch, V., Rader, T., Lyddiatt, A., Clerehan, R., Christensen, R., Boonen, A., Goel, N., et al. (2015). Toward ensuring health equity: Readability and cultural equivalence of omeract patient-reported outcome measures. *The Journal of rheumatology*, 42(12):2448–2459.

[Saggion et al., 2015] Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarevic, B. (2015). Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

[Spaulding, 1956] Spaulding, S. (1956). A spanish readability formula. *The Modern Language Journal*, 40(8):433–441.

[Štajner et al., 2015] Štajner, S., Mitkov, R., and Pastor, G. C. (2015). Simple or not simple? a readability question. In *Language Production, Cognition, and the Lexicon*, pages 379–398. Springer.

[Štajner and Saggion, 2013] Štajner, S. and Saggion, H. (2013). Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), Nagoya, Japan*, pages 374–382.

[Weiss et al., 2005] Weiss, B. D., Mays, M. Z., Martz, W., Castro, K. M., DeWalt, D. A., Pignone, M. P., Mockbee, J., and Hale, F. A. (2005). Quick assessment of literacy in primary care: the newest vital sign. *The Annals of Family Medicine*, 3(6):514–522.