Exploring social features for readability prediction

Ion Madrazo Azpiazu

1 Introduction

Reading is an important skill in the academic environment, a skill that can be critical for students' educational opportunities and their careers [19]. Giving students texts that suppose an increasing challenge to read during their learning process is essential. Even outside the educational environment, reading plays a important role. It is critical for people to fully comprehend the texts they read, specially when they face medical or legal issues. Understanding a legal or medical document properly, can lead the reader to make a better and more confident decision. However, studies [12, 16, 18] show that even medical documents that are supposed to be suited for average readers, tend to be too specialized and even well-educated adults have trouble understanding. Providing people with ways to ensure that the produced texts are simple enough for people with low reading abilities is imperative.

The readability or complexity of a text, and thus, the audience of it, can be assessed using a readability score. A readability score refers to the degree of ease with which a reader can understand a given text, a score which is usually determined by a readability formula. Historically, teachers have been the main stakeholders of readability formulas,using them for selecting new materials for their courses and curriculum design. However, lately, readability scores have been known to have more applicability than the ones in academic environments. Automatic text simplification [20,21], summarization for people with reading difficulties [7], book recommendation [17], literacy assessment [22], or even legal [13] and medical document complexity assessment [12, 16, 18] are only a few examples of applications that take advantage of the comprehension levels generated by existing readability scores.

Traditional formulas such as Flesh [10], Dale-Chall [4], and Gunning FOG [1] became very popular in the late 40's among the educators for manually determining text difficulty. Most of those formulas contained *shallow features* which provided a simple way of determining a texts complexity. However, they lacked precision in some cases, such as the one claimed in [6] where nonsense text could be classified as simple to read, just because it contained short and frequently used words. This encouraged researchers to study and develop better and more complex methods of prediction [2,11], that depended upon natural language processing and supervised machine learning techniques. These new formulas usually continued using the aforementioned shallow features, but added more complex features based on contents of the text.

Once the supervised learning framework became common in the field, most of the works have been focused on feature engineering. Some systems focused at lexical level [5, 9], using term based metrics such as frequencies, Tf-IDf or mutual information. However, these formulas were shown not to be domain independent [5], due to the fact that the features learned where to close to the topics of the text. Other systems [3, 8], focused their work on the structure of the texts

using syntactic features for capturing the complexity of the them. Semantic level features were also included in some readability assessment systems, such as the feature based on ambiguity the authors of [2] presented.

2 Project proposal

Readability assessment can be used in more than just plain text. Internet is evolving into a new social era and so are text resources too. Increasingly more resources contain social information such as hashtags, mentions or links, an information that is usually ignored by readability formulas. Therefore we would like to perform a preliminary research in order to see how the aforementioned information can be used for readability prediction, exploring with this, how social aspects of a texts can influence the readability of it. Furthermore, social media resources are known to be difficult to tackle by traditional natural language processing techniques, both because their lack of correctness and irregularity in terms of length. Therefore, we would also like to evaluate the performance traditional features have in social media texts.

3 Proposed method

The proposed system will be based on a supervised learning strategy. Given the ordinal and discrete nature of the class, we think that the task can fit both a regression and classification model. Therefore, both strategies will be tested during the development of the system. However, the main focus of the research will be on feature engineering, studying features that make use of social aspects in the texts i.e. mentions, hashtags, links or emoticons.

3.1 Text processing

Different text processing methods have been identified for the development of the system. Freeling NLP [14, 15] is a multilingual natural language processing (NLP) toolkit that supports 11 different languages. This tool solves common NLP tasks such as, tokenization, sentence detection, part of speech tagging or dependency parsing. WordNet is a lexical database that takes advantage of semantic relations between terms to build a graph that is very convenient for semantic analysis tasks. Latent semantic analysis is also a commonly used strategy for semantic analysis, which takes advantage of concurrences among terms for determining similarities between them. All those tools and others that we will discover in the process of development, will form the text processing step of the system.

3.2 Dataset

For development and evaluation purposes a dataset will be needed. The ideal dataset would be one containing social media documents with a level label attached to them. However, to the best of our knowledge, such dataset does not currently exist. For this reason, an approximation to the aforementioned ideal dataset will be created for the project. For doing so, we will analyze a twitter sample for the year 2015 consisting of 2% of the tweets created in that period. A graph will be generated using the mentions among the tweets and communities will be detected on it, using a yet to be determined algorithm. Those communities will be manually analyzed and tagged according to the readability level of the tweets the users inside them create. The tweets will be assigned the readability level of the community, generating with this a dataset

of leveled tweets.

References

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

[2] S. Aluisio, L. Specia, C. Gasperin, and C. Scarton. Readability assessment for text simplification. In *NAACL HLT*, pages 1–9, 2010.

[3] S. B. Bonsall, A. J. Leone, and B. P. Miller. A plain english measure of financial reporting readability. *Available at SSRN 2560644*, 2015.

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

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

[6] A. Davison and R. N. Kantor. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading research quarterly*, pages 187–209, 1982.

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

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

[9] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *COLING*, pages 276–284. Association for Computational Linguistics, 2010.

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

[11] T. François and C. Fairon. An ai readability formula for french as a foreign language. In *ACL EMNLP*, pages 466–477, 2012.

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

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

[14] L. PadrÃş, M. Collado, S. Reese, M. Lloberes, and I. CastellÃşn. Freeling 2.1: Five years of open-source language processing tools. In *LREC*, La Valletta, Malta, May 2010.

[15] L. PadrÃş and E. Stanilovsky. Freeling 3.0: Towards wider multilinguality. In *LREC*, Istanbul, Turkey, May 2012. ELRA.

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

[17] M. S. Pera and Y.-K. Ng. Automating readers' advisory to make book recommendations for k-12 readers. In *RecSys*, pages 9–16. ACM, 2014.

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

[19] R. D. Robinson, M. C. McKenna, and J. M. Wedman. Issues and trends in literacy education. 2000.

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

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

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