Rolling out Text Categorization for Language Learning Assessment supported by Language Technology

António Branco, João Rodrigues Francisco Costa João Ricardo Silva (1) and Rui Vaz (2)
(1) *Universidade de Lisboa*
Departamento de Informática, Faculdade de Ciências
(2) *Camões IP*
Divisão de Programação, Formação e Certificação

Abstract. This paper is concerned with a tool that supports human experts in their task of classifying text excerpts suitable to be used in quizzes for learning materials and as items of exams that are aimed at assessing and certifying the language level of students taking courses of Portuguese as a second language.

We assess the performance of this tool, which is currently available as an online service and is being used by the experts of the team of Camões IP that is responsible for the elaboration of the quizzes and of exam items to be used in language certification exams.

Keywords: language learning assessment, readability assessment, text categorization, second language skills certification, Portuguese language.

1 Introduction

As the Portuguese official language institute, Camões IP is responsible for running worldwide massive language learning courses of Portuguese as second language and for performing language certification exams. These courses and certification use quizzes and exams that resort to text excerpts belonging to one of the five language levels of skill A1, A2, B1, B2 and C1. As these excerpts can be used only once, specially in exams, the massive nature of these courses imposes a continuous task of always selecting more excerpts to be used in the upcoming exams and quizzes.

To help cope with this stringent demand, and support the instructors in their selection of these excerpts, an online tool was developed that seeks to help assigning each input excerpt to a language level. It returns the scores of a few metrics calculated with the use of natural language processing techniques. The design options and further details of this tool are presented and discussed in [1].

As this tool was made available as an online service,[1] we called LX-CEFR, it became nevertheless useful both for instructors and students. It is expected to help to improve the productivity and the consistency of the classification of candidate excerpts

[1] `http://lxcefr.di.fc.ul.pt`

by the experts. And it helps also to enhance the level of interactivity of the language courses for students, as they can resort to this tool to check whether a new text they come across may be appropriate for their current language level.

In this paper we report on a first evaluation exercise we performed over this tool, including an assessment of how inherently difficult is the task (for humans) and thus what may be the upper bound for the performance of a tool like this.

2 Language levels and metrics

Following CEFR[2] [2], Camões IP's language courses and language proficiency certification encompass five levels, of increasing difficulty: A1, A2, B1, B2 and C1. In order to help the user decide for the categorization of the input texts into one of these levels, the online tool displays the scores for fifteen metrics, covering for instance, number and proportion of letters, syllables, and words; number and proportion of simple, passive and subordinate clauses; number and proportion of nouns, verbs, prepositions, etc.; proportion of word types with only one occurrence; etc.

Out of these metrics, four are singled out as primary metrics and their scores offered in a radar chart (Figure 1). They are metrics that in the literature have been argued to correlate well with levels of readability [3]:

- Flesch Reading Ease index [4]
- Lexical category density in proportion of nouns
- Average word length in number of syllables per word
- Average sentence length in number of words per sentence

The Flesch metric is a wider accepted and used readability metric, which combines the third and fourth metric. Its higher scores indicate texts that are easier to read.

3 Language processing and projection of scores

To obtain the scores for these metrics, the online tool resorts to a number of state of the art natural language processing tools for Portuguese, such as tokenizers, POS taggers and syntactic parsers [5][6][7]. While the scores of the eleven secondary metrics are delivered as they are obtained, the scores of the four primary metrics are further processed in order to be projected into a continuous linear scale with five major points, each corresponding to a language level.

The dataset available to support the development of this tool consisted of 125 text excerpts previously used in past exams, each classified into one of the five language levels, and all in all containing 690 sentences, with 12,231 tokens. The average scores for the four primary metrics were calculated for each one of the five language levels. For each such metric, a linear regression over those averaged values was obtained, and each language level projected into the regression line in the chart. This formed the basis for the reference scales, represented in the arms of the radar chart.

[2] Common European Framework of Reference for Languages: Learning, Teaching, Assessment

When a text is input to the online tool, its scores for these four metrics are calculated and projected into the radar chart according to the comparison with the reference values obtained with the 125 excerpt reference dataset just described above.

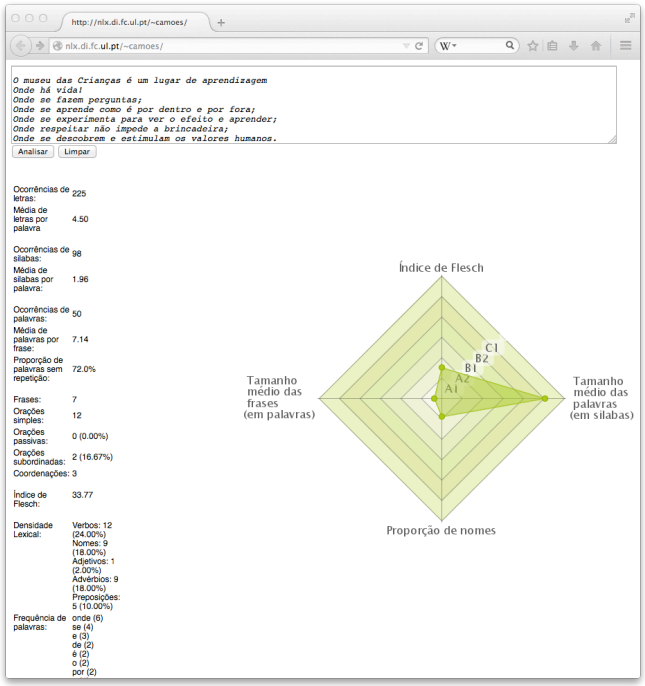

Figure 1: Online tool partial window with an example

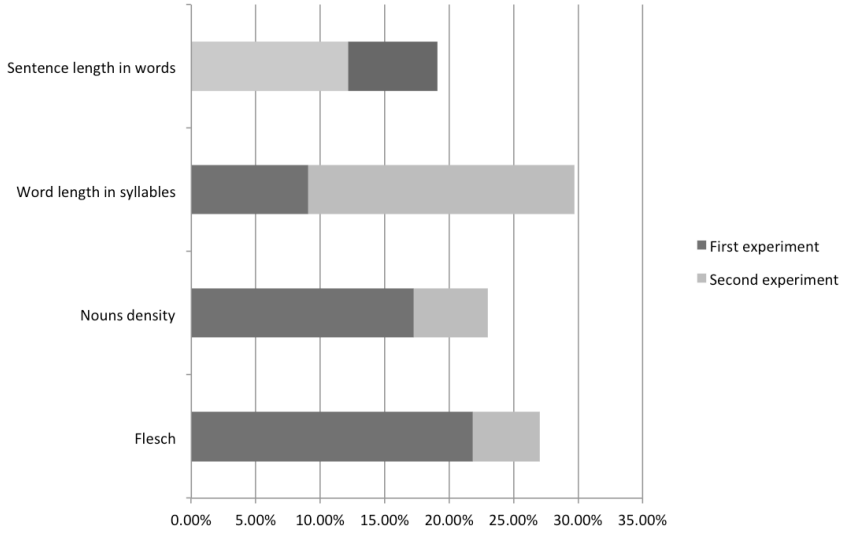

Figure 2: Accuracy in first and second experiment

4 Evaluation of the tool

To evaluate the tool we did a first evaluation experiment. We used this reference dataset and performed a 10-fold cross evaluation for each of the four primary metrics, obtaining accuracy values ranging from 9.00% (for word length) to 21.82% (Flesch index). The results are displayed in part of Figure 2 (in dark grey columns).

5 Assessment of the task

In order to put these scores into perspective, it was important to assess how inherently difficult could the task be in itself, that is how well human annotators executing it could perform.

To that end, the texts in the dataset were untagged of their originally assigned level, and five language instructors were recruited, which are trained and experts in selecting and classifying texts according to the relevant five CEFR levels. These experts performed the task of classifying each one of the texts. This re-annotated dataset permitted to assess the difficulty of the task along two measures: proportion of texts upon which there is agreement among annotators in their classification; inter-annotator agreement (ITA) given by Fleiss' kappa coefficient [8].

The distribution of the classifications of the annotators per CEFR level is displayed in the chart of Figure 3.

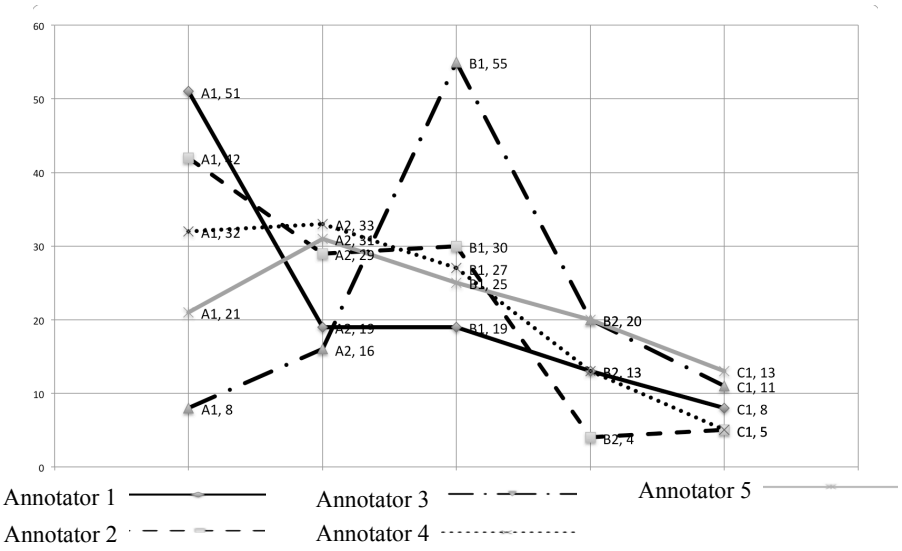

Figure 3: Number of texts assigned to each category (A1, A2, B1, B2 and C1) by each of the five annotators (Annotator 1 to 5)

The texts that received unanimous classification were 0.90% (only one text); those receiving the same classification by a majority of at least 4 classifiers were 17.27%. There were 67.27% of the texts receiving a given classification by a majority of at least 3 annotators.

The Fleiss' kappa coefficient value obtained for ITA was 0.13, corresponding to "Slight agreement", according to [9]. This is the second worst of five levels of agreement, and very distant from the value of 0.8+ widely assumed to be the level ensuring reliability of an annotated dataset.

6 Reevaluation of the tool

From the new reannotated dataset, we kept the 84 texts that received its classification by a majority of at least three annotators. With this subdataset, and with this new classification by several human experts, the reference scales for the four primary metrics were redone, following the same process as described above in the Section 3.

The evaluation of the ranking tool was redone, again with a 10-fold cross evaluation of the four primary metrics. The values now obtained after this fine tuning of the tool with the dataset annotated by multiple experts range from 12.16% (for sentence length) to 29.73% (word length), as displayed in part of Figure 2 (in light grey columns). There was a substantial improvement of three of the metrics with respect to the first evaluation, when the system was tuned with the dataset of texts used in previous exams, where each one was classified only by a single instructor.

7 Concluding remarks and future work

The results of the assessment of the task reveal that this is a quite difficult task even for humans, and hence represent a challenge for an automated ranking tool, which nevertheless is already attaining around 1/3 of its upper bound, as this is determined by the performance of humans.

It should be noted however that the volume of the dataset we could resort to may be too small and better results may eventually be expected with a larger dataset. From the first to the second experiment, all metrics improved except one, which may be deemed to the small size of the working sample.

Also worth noting is that, with a larger enough dataset, the superposition of a linear regression upon the distribution of the CEFR levels can be challenged and enhanced, and above all, advanced machine learning methods can be eventually benefited from.

Finally, also as future work, it will be important to undertake a usability assessment in order to gain a better understanding of how the tool can possibly be better adjusted to fit the needs of the human operators.

References

[1] Branco, António, João Rodrigues, Francisco Costa, João Ricardo Silva and Rui Vaz, forth., Text Classification for Interactive Language Learning.

[2] Council of Europe, 2011, consulted on http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp March 31, 2014.

[3] DuBay, William, 2004, *The Principles of Readability*, Costa Mesa, Impact Information.

[4] Flesch, Rudolf. 1979. *How to write in plain English: A book for lawyers and consumers*. New York: Harper.

[5] Branco, António and João Ricardo Silva, 2004, "Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese". In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (LREC2004), Paris, ELRA, pp.507-510.

[6] Branco, António and João Ricardo Silva, 2006, "LX-Suite: Shallow Processing Tools for Portuguese", In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL2006), Trento, Italy, pp.179-182.

[7] Silva, João Ricardo, António Branco, Sérgio Castro, and Ruben Reis 2010, "Out-of-the-Box Robust Parsing of Portuguese", In *Lecture Notes in Artificial Intelligence*, 6001, pp.86-89, Berlin: Springer.

[8] Fleiss, Joseph L., 1981, *Statistical methods for rates and proportions*. 2nd ed., New York: John Wiley, pp. 38–46.

[9] Landis, J. Richard, & Koch, Gary G., 1977, The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.