Hulat - ALexS CWI Task - CWI for Language and Learning Disabilities Applied to University Educational Texts

Rodrigo Alarcon[a], Lourdes Moreno[a] and Paloma Martínez[a]

[a] *Computer Science Department, Universidad Carlos III de Madrid, Leganés, Madrid, Spain*

Abstract

The number of citizens who face difficulties in reading and understanding written texts is growing. One of the possible cognitive accessibility barriers for cognitive, language and learning disabilities is when the texts contain unusual words. In this sense, there are a range of techniques that can be used to deal with this issue. Complex Word Identification (CWI), which aims to identify unusual words for a target audience, is one such technique. In this paper, a supervised architecture is described for the identification of complex words in university educational texts provided by the ALexS workshop. This architecture is composed of a Linear SVM with context-aware embedding features, provided by a BERT model. Moreover, easy-to-read and plain language resources were used. Our system participated in the ALexS CWI task, obtaining the second-best recall mark of 67%. However, low precision was due to, according to the analysis performed, having been trained with resources aimed at improving cognitive accessibility regardless of the domain. The results indicate that the level of readability and understanding is more demanding in informative fields, such as Wikipedia pages, than in the specific domain of university educational texts.

Keywords 1

Lexical simplification, CWI, Easy to read, BERT

1. Introduction

In the current era of information technology, information is abundant (education, news, social, health, government, etc.) for individuals. However, this information is not accessible to all people. Certain individuals face accessibility barriers when reading texts that contain long sentences, unusual words, complex linguistic structures, etc. Although people with intellectual and learning disabilities are most directly affected, cognitive accessibility barriers affect other user groups such as the deaf, deaf-blind, elderly, illiterate and immigrants with a different native language [1] [2]. People with reading disabilities can be found even among highly-educated users with specialized knowledge of the subject matter, such as university students. It may be possible to accommodate these users by making texts more readable.

In order to provide universal access to information and make texts more accessible, certain resources exist which provide helpful documentation, such as Easy-to-Read and plain language guidelines [3]. However, systematic compliance with these guidelines is complicated, and simplification processes, thus, become essential. Simplified versions are normally created manually. Manual simplification of written documents is quite expensive, particularly considering that information is continually being produced.

As a solution, Natural Language Processing (NLP) methods, such as text simplification, have been developed to provide systematic support and promote compliance with these cognitive accessibility guidelines, improving the readability and understandability of texts. There is a myriad of approaches to accomplish this goal, one being Complex Word Identification (CWI), which aims to identify words that are perceived as difficult for a given target audience.

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: ralarcon@inf.uc3m.es (A. 1); lmoreno@inf.uc3m.es (A. 2); pmf@inf.uc3m.es (A. 3)
ORCID: 0000-0003-3296-6873 (A. 1); 0000-0002-9021-2546 (A. 2); 0000-0003-3013-3771 (A. 3)

Considering this, a supervised CWI approach in the ALexS workshop is proposed in this paper which aims to identify complex words in university educational texts. The remainder of the paper is organized as follows. Section 2 briefly describes our training/test dataset and the ALexS dataset. In section 3, our system is described. Section 4 presents the task results obtained by our system, both with regards to the training/test stage and in the ALexS task. Finally, Section 5 offers conclusions.

2. Datasets

In order to follow a supervised approach, annotated data are necessary to identify whether a word is complex or simple. Therefore, our system was trained and tested with the following dataset.

2.1. Training/Test data

The data used was the annotated corpus of Spanish Wikipedia pages proposed in the BEA Workshop 2018 for the Complex Word Identification (CWI) (google.com/view/cwisharedtask2018) task. As shown in Table 1, 17603 instances were annotated by 54 Spanish speakers, most of whom were native [4].

Table 1.
Spanish CWI datasets distribution

	# Instances	# Complex	# Simple	# Uniwords
Training set	13748	5455	8293	11931
Development set	1622	653	969	1408
Test set	2233	907	1326	1955

Each instance contains a target uniword/multiword which is selected by annotators. Said target is marked as complex if at least one annotator designates it as complex. Moreover, each instance is represented by 11 columns which provide a range of different information. The dataset contains information for binary and probabilistic subtasks. For the development of this system, we focus on the binary classification subtask and we use the following information:

- **The Second Column** shows the actual sentence where a complex phrase annotation exists.
- **The Third Column** shows the start of the target word in the sentence.
- **The Fourth Column** shows the end of the target word in the sentence.
- **The Fifth Column** shows the target word.
- **The Tenth Column** shows the gold-standard label for the binary task (0: simple and 1: complex).

2.2. ALeXs Dataset

As shown in Table 2, the VYTEDU-CW corpus provided by the ALexS workshop (alexs-sepln-2020.org/) consists of 55 text files containing the video transcripts of classes given at the University of Guayaquil (Ecuador), resulting in a corpus of more than 68000 words, with more than 1200 words per transcription on average and 723 words which were designated as complex.

Table 2.
Spanish CWI datasets distribution

	Number of words	Number of Paragraphs
Min	465	5
Max	2646	18
Average	1241	907
Total	68248	613

3. Methods and system description

A supervised approach was proposed which aimed to identify complex words in educational texts.

3.1. Pre-processing

The VYTEDU-CW corpus texts described were pre-processed following a series of steps. First, the texts were split into sentences and tokens using Spacy (www.spacy.io/), an opensource library that provides support for texts in different languages, including Spanish. Finally, these tokens are filtered according to the following POS tags:

- **ADJ:** Adjective
- **ADV:** Adverb
- **NOUN:** Noun
- **PROPN:** Proper noun

The filtered text was then converted into the same format as that used during the training stage, preparing it for the next step of the process.

3.2. Supervised classification approach

To process the text from the previous stage, a supervised approach was followed by training an SVM algorithm due to its successful performance in text classification tasks. Moreover, SVM was also one of the most used algorithms for this task in SemEval2016 [5]. Specifically, a Linear SVC was chosen as it is much faster [6], takes advantage of the fact that SVM has shown good performance in classifying sparse instances [7] and, finally, had better results than previous tests carried out with a different type of kernel [8].

Using the dataset described in section 3 and in order to train the algorithm, each word (instance) needed to be represented as a set of features to help distinguish between complex and simple words. The proposed features used are described below:

- **Length feature:** word length
- **Boolean feature:** if a word is composed of capital letters
- **E2R feature:** a new feature established by creating an Easy-to-Read (E2R) dictionary.
- **Word2vec feature:** pre-trained Word2Vec model vectors.
- **BERT feature:** Pytorch pre-trained BERT model vectors.

In relation to the E2R feature, we proposed a new feature by creating an E2R dictionary that follows E2R guidelines. The goal of this feature was to optimize the detection of simple words. If a target word exists in the E2R dictionary, it receives a 0, otherwise is marked with a 1. The dictionary is fed from different sources that provide E2R texts drafted by experts with the support of the “Plena Inclusion” organization (/www.plenainclusion.org/). Some of these sources were: the Noticias fácil news page (www.noticiasfacil.es/) and the Easy Reading Association (www.lecturafacil.net/es/). Subsequently, this text was "cleaned" in order to preserve only the content words (noun, verbs, adjectives, adverbs). Currently, this dictionary contains 13400 simple words.

In the Word2vec feature, supported by the genism library, vectors were extracted for each word from a 300-dimension Word2vec model trained on the Spanish Billion Words Corpus [9].

The BERT feature operated in the following manner. Vectors were extracted for each word from a BERT (Bidirectional Encoder Representations from Transformers) [10] model (www.github.com/shehzaadzd/pytorch-pretrained-BERT). In order to do this, a 12-layer multilingual BERT pre-trained model was used first before word vectors were extracted by adding the last four layers and using the first 480 dimensions of the model. To do this first we use the stored hidden states of the model that has four dimensions: the layer number, the batch number (one sentence per instance), the token number and the feature number (768 features). Later, our word vectors for each word of the sentence are created by summing the last four layers. These layers are selected because they've shown better results in our tests and it can show different results depending on the task.

These embeddings are useful for semantic searches and information retrieval. The main difference between this type of embedding and others, such as Word2Vec or FastText, is that BERT produces word representations that are dynamically informed by the words around them, whereas Word2Vec the words are represented as unique indexed values. In the common word embedding models, each word is represented with one single vector, ignoring polysemy words. In a sense, with word embedding, each word could have several vectors, one for each of its possible meanings. Therefore, these models allow us to deal with the task of word disambiguation when we identify complex words.

4. Results

Due to the fact that a validation dataset from the workshop was not given, the BEA's workshop test dataset was used to validate our features and make adjustments. Table 3 shows the results obtained as regards the Train and Train+Developer datasets, which were validated with the test dataset. The results in both cases outperformed the results obtained by other systems from the abovementioned workshop [11][12][13][14]. Subsequently, a model was trained with the Train+Developer+Test to process the ALexS dataset, which, when evaluated with the test dataset, obtained a result of 0.81.

Table 3.
Results for the test dataset

	Accuracy	Precision	Recall	F1 Score
TRAIN	0.80	0.79	0.78	0.792
TRAIN+DEV	0.80	0.80	0.79	0.794

Additionally, to complement previous information, Table 4 shows the scores of some combinations between these features, helping us determine which features are more discriminatory. One of the best scores are reached with the help of vectors of the embedding models. Using Word2Vec and BERT models, a F1-score of 0.752 is obtained. Also, evaluating F1-scores independently for each feature, BERT feature shows a F1-score of 0.727, being the best score between all independent features. Likewise, the W2V feature yields a score of 0.70, proving to be a valuable resource for this task.

Table 4.
Results for the test dataset

	Accuracy	Precision	Recall	F1 Score
L+BT	0.79	0.78	0.77	0.778
L+B+BT	0.79	0.79	0.78	0.783
L+B+E+BT	0.80	0.80	0.78	0.787
L+B+E+W+BT	0.80	0.80	0.79	**0.794**
W+BT	0.77	0.76	0.75	**0.752**
L	0.73	0.74	0.70	0.702
BT	0.74	0.74	0.72	**0.727**
W	0.72	0.71	0.70	0.700

Table 5 shows the results of the CWI task in ALexS workshop. Although it gives a lower precision (with a score of 0.9), the system received the second-highest rank in Recall (with a score of 0.67), obtaining good coverage on the detection of words. The generalization of the system seems to be good. However, it needs to improve on specific domains when dealing with technical words.

Table 5.
Results on ALexS task

Participants	Accuracy	Precision	Recall	F1 Score
Antonio Rico - Method 1	0.98	0.33	0.22	0.26
Antonio Rico - Method 2	0.98	0.34	0.23	0.27
Antonio Rico - Method 3	0.98	0.33	0.22	0.26
Elena Zotova - Method 1	0.91	0.10	0.60	0.17
Elena Zotova - Method 2	0.89	0.09	0.69	0.16
Elena Zotova - Method 3	0.91	0.10	0.59	0.17
George Zaharia	0.91	0.02	0.08	0.03
(*) Rodrigo Alarcón (HULAT)	0.90	0.09	0.67	0.16
AlexS 2020 Organizers	0.92	0.12	0.66	0.20

In addition to the previous information, Table 6 confirms the previously described information and shows the results on some of the texts of the VYTEDU-CW corpus. For example, in the text of video 41, the system showed good recall on the task by predicting all the complex words. However, at the same time, it presented several false positives due to the generalization issue.

To illustrate this, take, for example, the word "biodiversidad" (*biodiversity*) that would be labeled as complex in a generic domain. Nevertheless, this same word, in a university educational domain, would be labeled as simple. This can be confirmed by comparing these annotations with the dataset from the BEA Workshop comprised of annotated Wikipedia pages which contain generic content. In this instance, the word "investigaciones" (*investigations*) is labeled as simple in the ALexS dataset but complex in the BEA dataset. There are many examples such as this in which the reason why the system shows good recall, but low precision is demonstrated.

Based on the outcomes, it can be seen that the university educational texts are not easily readable for university students with language and learning disabilities.

Table 6.
Specific results on the ALexS task (Participant: Rodrigo Alarcón)

	Accuracy	Precision	Recall	F1 Score
Video 5	0.92	0.56	0.43	0.49
Video 41	0.92	0.05	**1**	0.10
Video 43	0.86	0.02	**1**	0.04
Video 48	0.97	0.04	**1**	0.08

5. Conclusions and future work

The main objective of this work is to improve cognitive accessibility by increasing the understanding and readability of texts. In order to accomplish this objective, a supervised algorithm that uses a more refined, context-aware embedding model and Easy-to-Read resources was trained. The experiments showed that the combinations of these features with a Linear SVM outperforms previous systems. However, it also presented difficulties when dealing with specific domains with less of a demand for readability, such as educational texts at a university level. To improve the precision of our system and a obtain a better result in the classification, university educational domain resources should be used. BERT and Word2Vec models with university educational texts can be trained. Additionally, students with language and learning disabilities should not be considered as the target audience however, at the university there are students with language and learning disabilities.

Regarding the approach followed in our complex word detection system, one of the main contributions of this research work has been the use of BERT embeddings in the prediction. For future work, we plan to explore more features of BERT models. With the extracted vectors, we can evaluate the cosine distance between the target word and the surroundings in the sentence. By giving this additional information, provided by more detailed embedding, a better score in the CWI task can be accomplished. At the same time, we can evaluate the synergy between a wider variety of embeddings, such as Sense2Vec [15] and Char2Vec.

6. Acknowledgements

This work was supported by the Research Program of the Ministry of Economy and Competitiveness-Government of Spain, (DeepEMR project TIN2017-87548-C2-1-R)

7. References

[1] D. Ferrés, H. Saggion, and X. G. Guinovart, “An adaptable lexical simplification architecture for major ibero-romance languages,” in Proceedings of the first workshop on building linguistically generalizable NLP systems, (2017), pp. 40–47.

[2] L. Moreno, P. Martínez, I. Segura-Bedmar, and R. Revert, “Exploring language technologies to provide support to WCAG 2.0 and E2R guidelines,” Proc. XVI Int. Conf. Hum. Comput. Interact. - Interacción ’15, pp. 1–8, (2015).

[3] L. Moreno, R. Alarcon, and P. Martínez, “Lexical simplification approach to support the accessibility guidelines,” Proceedings of the XX International Conference on Human Computer Interaction (Interacción ’19), pp. 1–4, (2019).

[4] S. M. Yimam, S. Stajner, M. Riedl, and C. Biemann, “Multilingual and Cross Lingual Complex Word Identification,” Recent Adv. Nat. Lang. Process., pp. 813–822, (2017).

[5] G. Paetzold and L. Specia, “SemEval 2016 Task 11: Complex Word Identification,” Proc. 10th Int. Work. Semant. Eval., pp. 560–569, (2016).

[6] I. Segura-bedmar, C. Colón-ruíz, M. Á. Tejedor-alonso, and M. Moro-moro, "Predicting of anaphylaxis in big data EMR by exploring machine learning approaches," J. Biomed. Inform., vol. 87, no. January, pp. 50–59, (2018).

[7] J. Kivinen, M. K. Warmuth, and P. Auerc, "The Perceptron algorithm versus Winnow : linear versus logarithmic mistake bounds when few input variables are relevant," vol. 97, no. 97, pp. 325–343, (1997).

[8] R. Alarcon, L. Moreno, I. Segura-bedmar, and P. Martínez, "Lexical simplification approach using easy-to-read resources," Proces. del Leng. Nat., pp. 95–102, (2019).

[9] C. Cardellino, "Spanish {B}illion {W}ords {C}orpus and {E}mbeddings." (2016).

[10] M. C. Kenton, L. Kristina, and J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," no. Mlm, 1953.

[11] S. M. Yimam et al., "A Report on the Complex Word Identification Shared Task 2018," (2018).

[12] D. De Hertog and K. U. Leuven, "Deep Learning Architecture for Complex Word Identification," pp. 328–334, (2018).

[13] D. Alfter, "SB @ GU at the Complex Word Identification 2018 Shared Task," pp. 315–321, (2018).

[14] T. Kajiwara and M. Komachi, "Complex Word Identification Based on Frequency in a Learner Corpus," pp. 195–199, (2018).

[15] A. Trask, P. Michalak, and J. Liu, "sense2vec - A Fast and Accurate Method for Word Sense Disambiguation In Neural Word Embeddings," pp. 1–9, (2015).