Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN.

Jenny A. Ortiz-Zambrano[a], Arturo Montejo-Ráez[b]

[a]Universidad de Guayaquil. Guayaquil, Ecuador
[b]CEATIC. Universidad de Jaén, Jaén, España

Abstract

In September 2020, the first edition of the ALexS workshop (Task on Lexical Analysis at SEPLN) was held in Málaga, Spain as part of the second edition of IberLEF (Iberian Languages Evaluation Forum), which joined the efforts of the IberEval and TASS workshops. In this first edition, there has been only one task proposed: Complex Word Identification (CWI). More than seven teams joined the campaign, but only three of them finally submitted results and a description of their systems. The difficulty of the task, due to the lack of labeled data to participants, has forced interesting approaches to tackle the CWI problem in a unsupervised o semi-supervised way. This paper summarizes the approaches and the results of the submitted systems by different teams.

Keywords

Lexical Analysis, Complex Word Identification, Text Simplification.

1. Introduction

Information is delivered under several forms depending on their content and audience: digital newspapers, academic publications, blogs... and, among them, educational texts. Students hold different levels in reading comprehension, being the main barrier the vocabulary present in the text: it is more important to understand words than grammar complexity in most cases. ALexS workshop proposes a Complex Word Identification (CWI) task on educational text at university level.

The ALexS (*Análisis Léxico en SEPLN 2020*) Lexical Analyis Workshp at SEPLN, was part of Iberian Languages Evaluation Forum (IberLEF) [1], hosted by and co-located with the International Conference of the Spanish Society on Natural Language Processing (SEPLN). IberLEF is the result of the association of some workshops in the Natural Language Processing (NLP) domain for Spanish and other languages spoken in the Iberian peninsula, and the aim of joining forces of different NLP research communities in order to provide a common forum for assessing NLP systems and interchanging research ideas, issues, challenges and experiences.

Although there are some tasks were Spanish is considered for Complex Word Identification, like the CWI Shared Tasks in SemEval 2016 [2] and NAACL-HTL 2018 [3], a new annotated

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: jenny.ortizz@ug.edu.ec (J.A. Ortiz-Zambrano); amontejo@ujaen.es (A. Montejo-Ráez)
URL: https://www.ujaen.es/ (A. Montejo-Ráez)
ORCID: 0000-0002-8643-2714 (A. Montejo-Ráez)

CEUR Workshop Proceedings (CEUR-WS.org)

corpus of transcriptions of teaching classes at the University of Guayaquil (Ecuador), the VYTEDU-CW corpus [4], was recently created. This resource can be used to test complex words identification systems, configured to fit in an educational scope.

Seven research teams submitted several classification results to the Subtask 1, and four teams submitted to the Subtask 2. The systems submitted go in the line of the state of the art in similar workshops, and the participants developed classification systems based on Recurrent Neural Networks, Transformer Networks and fine-tunning models built upon BERT [?]. The details of the systems submitted are described in Sections ?? and ??.

2. Complex Word Identification Task

Complex word identification (CWI) is a common task within the more general of text simplification [5]. Actually, in order to perform lexical simplification, words considered difficult to the reader have to be identified first. It is of interest in areas like second language acquisition [6] or reading comprehension [7] for people with certain disabilities. As stated before, only two shared tasks have tackled the problem in recent years, the CWI Shared Tasks at SemEval 2016 [2] and NAACL-HTL 2018 [3]. The aim of these task was to contribute in the advance of methods and techniques for effective complex word identification, as the substitution of complex words in texts improves the understandability of a given text by the reader (thus, exhibiting a better readability level). Many areas are interested in this task, as stated before, and still active research is being performed towards such a goal.

The objective is to mark those words that can be considered complex, in the sense of difficult comprehension for the reader. The corpus used in this workshop is the VYTEDU-CW corpus. There are some interesting challenges in this task compared to other CWI tasks:

- Difficult terms have to be within the scope of an academic content. That is, many technical terms may need to be superseded as they are commonly used in the domain.
- There are several domains corresponding to different grades, so the system has to adapt to them.
- No training data will be released, only dev data for adjusting systems to file formats. Therefore, non-supervised or semi-supervised approaches are applicable.

Next, more details on the VYTEDU-CW corpus are given.

2.1. The VYTEDU-CW corpus

An adhoc corpus has been created, as a variant of the original VYTEDU-CW corpus[4]. The collection contains 55 texts which correspond to transcripts of academic videos in Spanish made within the classrooms of the different careers of the University of Guayaquil. The VYTEDU-CW (Videos and transcripts in the educational field - Complex words) corpus is conformed by more than 1,200 words per transcription on average, has a total of 9,175 different words. Its data set contains 723 words annotated as complex (difficult) terms that are present in the different documents, these difficult words were identified and labeled by 430 annotators (students), 250 students tagged words that other users had not selected, that is, that did not match those annotated by other students.

The data set that makes up VYTEDU-CW are seven fields:

- The word identified and labelled as complex.
- The student's identification.
- The name of the document read.
- The initial position in the text of the difficult word,
- The length of the word in characters,
- The date and time of the creation of the annotation.

Some examples of the words labeled by the students correspond to names of characters, abbreviations, use of sophisticated terminology by teachers when teaching their classes, the use of technical words by teachers, the use of nominal verbs, another problem that exists is that teachers propose examples in class using words that do not belong to the level of education or specialization, the use of long words, teachers use words difficult to pronounce and unusual, the use of compound words, the use of proverbs to illustrate examples, among others.

We can observe an example in the use of long words, or of difficult pronunciation that caused the students difficulty in being able to identify them, such was the case in the career of ”Law”, the students labeled the words, such as: interculturality, methodological, homogeneity.

The use of abbreviations is also the cause of the barriers that are formed in the understanding of students. For example, in the ”Networking” degree the abbreviations were labeled: LAN, GEAR IPAN, MBPS. Another word identified by the students as difficult was in the ”Research Unit”, the abbreviation ISBN.

2.2. The task: complex word identification

The task proposes the tagging of words (or multi-words) from lecture transcriptions that could be considered difficult to understand for students. Annotations in the corpus were made by students at the same academic level of the annotated lecture. Therefore, we are facing an scenario different from that researchers on CWI are used to. Besides, no training neither development sets have been released, only the number of total annotations in the corpus is revealed to participants. This task encourages the exploration of unsupervised approaches to complex word identification.

As this is a first tentative to academic/educational CWI, measure will be classical Precision, Recall and F-score.

3. Systems presented

Three teams presented their systems and results for this first task. Their approaches are summarized below.

3.1. UDLAP participation

The UDLAP team is composed of only one member: Antonio Rico-Sulayes, from Universidad de las Américas in Mexico. The approach proposed is a pipeline of several filters [8]:

1. A general lexicon (CREA [9]) that has been extended with proper names and verb conjugates.
2. An specialized lexicon of Internet-related terms.
3. A filter on frequent n-grams (extracted from the general lexicon)
4. A filter based on normalized frequency over corpus documents

The thresholds and parameters over each module were selected to produced a final expected quantity of complex-word candidates as close as possible to that exhibited by the corpus, as this number is known by participants.

This system reached the highest score on the macro F1 metric, the highest macro precision and the highest recall over all runs submitted to ALexS by different participants.

3.2. Vicomtech participation

The Vicomtech team is composed of for members, belonging to the SNLT group at Vicomtech Foundation, from San Sebastián, Spain.

The approach followed by these participants [10] is the only one that takes into account the different categories or domains that can be identified in the VYTEDU-CW corpus, which are related to several academic learning profiles (degrees) wherefrom the texts were generated. Each word is modeled as seven different word-level features (lemma length, lemma frequency in subject documents, number of synsets in WordNet, lemma frequency in domain corpora, lemma probability in domain corpora, word frequency in Wikipedia and word probability in Wikipedia). Then, within each domain, a clustering process is done. The parameters of this unsupervised step have been taken from other complex-words detection tasks. Finally, for each domain, different groups of words are obtained, considering as complex that group with lowest average value of features. K-Means with the seven features showed the best Recall, Accuracy and G-score (harmonic mean between Accuracy and Recall).

3.3. HULAT participation

The third team participating in ALexS 2020 is another Spanish group of researchers. Its three authors are from the Computer Science Department of Universidad Carlos III (Madrid, Spain).

The solution proposed for solving the CWI task is a supervised one [11], with an architecture enconding each word using different features: word length, a boolean determining whether only capital letters are used, a boolean determining its inclusion in an easy-to-read lexicon, Word2Vec vectors and BERT vectors (these last two using pretrained models).

Finally, all these features are concatenated and fed into a SVM classifier, which was trained and fine-tuned using the Spanish partition of the BEA Workshop 2018 CWI task dataset.

Table 1
ALexS - Official results

Team	Run	Macro F1	Macro Precision	Macro Recall
UDLAP	Method 2	**0.272528**	**0.341598**	0.226691
UDLAP	Method 1	0.266004	0.334722	0.220696
UDLAP	Method 3	0.261395	0.328729	0.216955
Vicomtech	Method 1	0.166623	0.096827	0.596863
Vicomtech	Method 3	0.176915	0.103961	0.593173
Vicomtech	Method 2	0.162051	0.091838	**0.688192**
HULAT	Method 1	0.164521	0.093712	0.673171

4. Results

Table 1 shows the results obtained on the ALexS task for complex word identification on the VYTEDU-CW dataset. UDLAP team obtained the overall best results in terms of F-score (0.2725) and precision (0.3415). Vicomtech system performed well on recall (0.6881), but at the price of a very low precision. Although up to three different runs were allowed, only the UDLAP and Vicomtech submitted three annotations, while HULAT team only submitted one annotation. As organizers, we expected more participants to submit their results, being more than 10 teams registered at the beginning of the campaign. This could be due to the COVID-19 pandemic that is affecting everything in our lifes. Anyhow, the systems proposed were different enough to extract interesting conclusions from such an small participation.

5. Conclusions

The results obtained were poor, although a significant recall value was obtained. The authors agree on the negative effect of training on a corpus and transferring the learned model to a different domain. It was clear that the corpus and the task was very challenging, as there was no clue about which type of words could be considered complex, the rate of them over the corpus or examples to have some insight. Anyhow, the description of the task stated that there were different domains, and that complex words to be identified should adapt to each domain. That is, a word like “stock” could be considered complex in History, but not in Economics.

Two of the system proposed generated several words characteristics, and were fed into supervised (SVM) o non-supervised (clustering) algorithms to determine whether the word was complex or not. The other system relied on pure lexicon-based filtering according to probability (i.e. frequency) of appearance. A promising solution could be a hybrid system with all those different characteristics (even the frequency/probability over different lexicons) all together in a non-supervised algorithm, which only needs a threshold parameter to fine-tune the identification of the class of complex words.

If this task finds its continuity for next year, it is expected to extend the corpus and provide a training subset to users. Also, following the design of the CWI task in SemEval 2021, we could consider isolated words versus multi-words as to different subtasks. Overall, more research on CWI for Spanish is needed in order to improve future text simplification systems.

Acknowledgments

This work has been partially supported by a grant from the Spanish Government under the LIVING-LANG project (RTI2018-094653-B-C21) . Eugenio Martínez Cámara was supported by the Spanish Government Programme Juan de la Cierva Formación (FJCI-2016-28353).

References

[1] IberLEF 2020 web site, http://sepln2020.sepln.org/index.php/iberlef/, ???? Accessed: 2020-06-30.

[2] G. Paetzold, L. Specia, SemEval 2016 task 11: Complex word identification, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California, 2016, pp. 560–569. URL: https://www.aclweb.org/anthology/S16-1085. doi:10.18653/v1/S16-1085.

[3] S. M. Yimam, C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, M. Zampieri, A report on the complex word identification shared task 2018, in: Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 66–78. URL: https://www.aclweb.org/anthology/W18-0507. doi:10.18653/v1/W18-0507.

[4] J. A. Ortiz-Zambrano, A. Montejo-Ráez, K. N. Lino-Castillo, O. R. González-Mendoza, B. C. Cañizales-Perdomo, VYTEDU-CW: Difficult words as a barrier in the reading comprehension of university students, Advances in Emerging Trends and Technologies: Volume 1 1066 (2019) 167.

[5] H. Saggion, Automatic text simplification, Synthesis Lectures on Human Language Technologies 10 (2017) 1–137.

[6] R.-M. Botarleanu, M. Dascalu, S. A. Crossley, D. S. McNamara, Sequence-to-sequence models for automated text simplification, in: I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, E. Millán (Eds.), Artificial Intelligence in Education, Springer International Publishing, Cham, 2020, pp. 31–36.

[7] P. Chen, J. Rochford, D. N. Kennedy, S. Djamasbi, P. Fay, W. Scott, Automatic text simplification for people with intellectual disabilities, Artificial Intelligence Science and Technology (2016).

[8] A. Rico-Sulayes, General lexicon-based complex word identification extended with stem n-grams and morphological engines, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 2020.

[9] Corpus de Referencia del Español Actual (CREA) - Listado de frecuencias, http://corpus.rae.es/lfrecuencias.html, ???? Accessed: 2020-07-30.

[10] E. Zotova, M. Cuadros, N. Perez, A. García-Pablos, Vicomtech at ALexS 2020: Unsupervised complex word identification based on domain frequency, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 2020.

[11] R. Alarcón, L. Moreno, P. Martínez, Hulat - ALexS CWI task - CWI for language and learning disabilities applied to university educational texts, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR-WS, Malaga, Spain, 2020.