Complex Word Identification:
Challenges in Data Annotation and System Performance

Marcos Zampieri1, Shervin Malmasi2, Gustavo Paetzold3, Lucia Specia3
1University of Wolverhampton, United Kingdom
2Harvard Medical School, United States
3University of Sheffield, United Kingdom

m.zampieri@wlv.ac.uk, smalmasi@bwh.harvard.edu
g.h.paetzold@sheffield.ac.uk, l.specia@sheffield.ac.uk

Abstract

This paper revisits the problem of complex word identification (CWI) following up the SemEval CWI shared task. We use ensemble classifiers to investigate how well computational methods can discriminate between complex and non-complex words. Furthermore, we analyze the classification performance to understand what makes lexical complexity challenging. Our findings show that most systems performed poorly on the SemEval CWI dataset, and one of the reasons for that is the way in which human annotation was performed.

1 Introduction

Lexical complexity plays a crucial role in reading comprehension. Several NLP systems have been developed to simplify texts to second language learners (Petersen and Ostendorf, 2007) and to native speakers with low literacy levels (Specia, 2010) and reading disabilities (Rello et al., 2013). Identifying which words are likely to be considered complex by a given target population is an important task in many text simplification pipelines called complex word identification (CWI). CWI has been addressed as a stand-alone task (Shardlow, 2013) and as part of studies in lexical and text simplification (Paetzold, 2016).

The recent SemEval 2016 Task 11 on Complex Word Identification – henceforth SemEval CWI – addressed this challenge by providing participants with a manually annotated dataset for this purpose (Paetzold and Specia, 2016a). In the SemEval CWI dataset, words in context were tagged as complex or non-complex, that is, difficult to be understood by non-native English speakers, or not. Participating teams used this dataset to train classifiers to predict lexical complexity assigning a label 0 to non-complex words and 1 to complex ones. Below is an example instance from their dataset:

(1) A frenulum is a small fold of tissue that secures or restricts the motion of a mobile organ in the body.

The words in bold — frenulum, restricts, motion — have been assigned by at least one of the annotators as complex and thus they were labeled as such in the training set. All words that have not been assigned by at least one annotator as complex have been labeled as non-complex.

In this paper we evaluate the dataset annotation and the performance of systems participating in the SemEval CWI task. We first estimate the theoretical upper bound performance of the task given the output of the SemEval systems. Secondly, we investigate whether human annotation correlates to the systems’ performance by carefully analyzing the samples of multiple annotators. Although in the shared task complexity was modeled as a binary classification task, we pose that lexical complexity should actually be seen in a continuum spectrum. Intuitively, words that are labeled as complex more often should be easier to be predicted by CWI systems. This hypothesis is investigated in Section 3.3. To the best of our knowledge, no evaluation of this kind has been carried out for CWI. The most similar analyses to ours have been carried out by Malmasi et al. (2015) for native language identification and by Goutte et al. (2016) for language variety identification.

2 Methods and Experiments

In this section we present the data, the methods, and an overview of the experiments we propose in this paper. The goal of the experiments is to evaluate CWI performance with respect to computational methods and the manual annotation of the
Table 1: SemEval CWI - Systems and approaches

| Team        | Approach                                                                 | System Paper                        |
|-------------|---------------------------------------------------------------------------|-------------------------------------|
| SV000gg     | System voting with threshold and machine learning-based classifiers trained on morphological, lexical, and semantic features | (Paetzold and Specia, 2016b)        |
| TALN        | Random forests of lexical, morphological, semantic & syntactic features   | (Ronzano et al., 2016)              |
| UWB         | Maximum Entropy classifiers trained over word occurrence counts on Wikipedia documents | (Konkol, 2016)                      |
| PLUJAGH     | Threshold-based methods trained on Simple Wikipedia                        | (Wróbel, 2016)                      |
| JUNLP       | Random Forest and Naive Bayes classifiers trained over semantic, lexicon-based, morphological and syntactic features | (Mukherjee et al., 2016)            |
| HMC         | Decision trees trained over lexical, semantic, syntactic and psycholinguistic features | (Quijada and Medero, 2016)          |
| MACSAAR     | Random Forest and SVM classifiers trained over Zipfian features            | (Zampieri et al., 2016)             |
| Pomona      | Threshold-based bagged classifiers with bootstrap re-sampling trained over word frequencies | (Kauchak, 2016)                    |
| Melbourne   | Weighted Random Forests trained on lexical/semantic features              | (Brooke et al., 2016)               |
| IIIT        | Nearest Centroid classifiers trained over semantic and morphological features | (Palakurthi and Mamidi, 2016)      |

The task was very popular, having attracted 21 teams and 42 participating systems. In Table 1 we present the 10 highest performing approaches proposed by participants of the SemEval CWI task.
of the top 10 systems. Our assumption was that including systems that did not perform well in the task degrades the voting performance by introducing too much noise in the predictions.

Plurality voting results for class 1 are presented in Table 2 in terms of precision, recall, and F1 score. For comparison we also report a threshold-based baseline on word frequencies from Wikipedia (Paetzold and Specia, 2016a) and the performance of the best system in terms of f-score for class 1. The number of instances in each class is presented in the column ‘Samples’.

| System  | Class | P     | R     | F1    | Samples |
|---------|-------|-------|-------|-------|---------|
| All     | 0     | 0.98  | 0.83  | 0.90  | 84,090  |
| All     | 1     | 0.17  | 0.71  | 0.27  | 4,131   |
| Top 10  | 0     | 0.98  | 0.88  | 0.93  | 84,090  |
| Top 10  | 1     | 0.21  | 0.66  | 0.32  | 4,131   |
| Baseline| 0     | 0.08  | 0.90  | 0.15  | 4,131   |
| Best    | 1     | 0.29  | 0.45  | 0.35  | 4,131   |

Table 2: Results for plurality voting

The results obtained show that the plurality voting system performs significantly better on class 0 (non-complex words) achieving 0.90 F1 score than on class 1 (complex words) achieving 0.27 F1 score. The majority of instances in the dataset are non-complex words and this explains the bias. For class 1, the F1 score obtained by the ensemble featuring the top 10 systems outperforms the baseline but it is outperformed by the best system by 3 percentage points.

3.2 Optimal Ensemble and Oracle

We showed the performance of plurality voting ensembles built with the output of all systems and with the output of the top-10 ranked systems. The setup using the output of the top-10 systems yielded very good performance, but still below the best system in the competition. In this section we investigate how many systems should be included in the ensemble to obtain the best possible performance. In Figure 1 we show the F1-score, precision, and recall results for class 1 obtained by plurality voting using ensemble configurations ranging from 3 to 46 systems.

To investigate the optimal ensemble configuration we performed a greedy backward search over the systems, iteratively removing the worst systems in a stepwise manner without a stopping criterion. The best performance for complex words was obtained using with the predictions of the top-3 systems achieving 0.35 F1-score. This is the best performing and smallest ensemble configuration confirming that the SemEval CWI is a very challenging task which led the vast majority of systems to perform so poorly that the plurality voting ensemble did not benefit from their predictions.

Finally, in Table 3 we present the results obtained by the oracle classifier using the top-3 systems, which yielded the best results in the plurality voting ensemble. The oracle performs very well when predicting non-complex words achieving 0.98 F1-score. The performance for complex words was substantially higher than the one obtained using the configurations of the plurality voting ensemble, reaching 0.60 F1-score and outperforming both the baseline and the best system. This is the theoretical upper bound of the task given the output of the systems that used this dataset.

| System  | Class | P     | R     | F1    | Samples |
|---------|-------|-------|-------|-------|---------|
| Oracle  | 0     | 0.98  | 0.98  | 0.98  | 84,090  |
| Oracle  | 1     | 0.59  | 0.61  | 0.60  | 4,131   |
| Baseline| 1     | 0.08  | 0.90  | 0.15  | 4,131   |
| Best    | 1     | 0.29  | 0.45  | 0.35  | 4,131   |

Table 3: Results for oracle classifier (top-3)

3.3 Lexical Complexity

In this section we investigate features of the dataset and annotation that influence the output of the classifiers using the training set and the results of the 10 best performing systems. We start by looking at an histogram of annotations of all complex words in the training data (Figure 2).
Among the 2,237 words in the training set, 706 were labeled as complex. The histogram shows the distribution of the annotation that ranged from 393 words labeled by 1 annotator as complex and only 5 words labeled by all 20 annotators as such.

Inspired by readability metrics (Kincaid et al., 1975), we looked at the average word length (AWL) of the words in the training set under the assumption that longer words tend to be more often perceived as complex. We divide the dataset in intervals according to the number of annotators that assigned each word as complex: 10-20, 1-9, and none. Results are presented in Table 4.

Table 4: Word length and complexity

| Class | Annotators | Words | AWL  |
|-------|------------|-------|------|
| 1     | 10-20      | 42    | 7.07 |
| 1     | 1-9        | 664   | 6.71 |
| 1     | 1-20       | 706   | 6.74 |
| 0     | 0          | 1,531 | 5.94 |

Finally, we investigate the interplay between annotation and system performance by analyzing the 38 words in the training data which were labeled as complex by at least half of the annotators. We 1) check the overlap of these words in the training and test sets; 2) verify how many overlapping words received the same label in the training and test sets; 3) compute the number of times humans annotated a given word as complex (0-20) and the number of top-10 systems that labeled the word as complex (0-10). We present the scores for the words that met these criteria in Table 5. For comparison we also present five randomly selected words labeled as complex by only one annotator which received the same label in the train and test sets.

Table 5: Annotation vs. prediction.

| Word     | Humans | Systems |
|----------|--------|---------|
| gharial  | 20     | 10      |
| khachkar| 17     | 10      |
| anoxic   | 14     | 10      |
| ubiquitous | 12    | 8       |
| rebuffed | 11     | 10      |
| took     | 1      | 0       |
| better   | 1      | 0       |
| however  | 1      | 0       |
| designation | 1     | 4       |
| islands  | 1      | 0       |

The CWI dataset replicates a scenario in which the vocabulary limitations of individuals is assessed based on the overall limitations of a group, as a result 50% of the most complex words did not receive the same label in the training and test sets. Nevertheless, the results of this pilot analysis seem to confirm our hypothesis that words that were tagged more often as complex in the training set tend to be easier for CWI system to identify.

4 Conclusion and Future Work

This paper complements the findings from the SemEval CWI shared task report (Paetzold and Specia, 2016a) by presenting an evaluation of CWI system outputs and of the dataset used in the shared task. We were able to: 1) estimate the potential upper limit of the task considering the output of the participating systems (0.60 F1 score for complex words); 2) provide empirical evidence of the relation between word length and lexical complexity for this dataset; and 3) confirm that the performance of CWI systems in this shared task is related to non-native speakers’ annotation.

Our findings serve as a starting point for a potential re-run of the SemEval CWI task and for other studies using the 2016 dataset. In future work we would like to investigate other factors that influence lexical complexity such as word frequency and grammatical categories.
Acknowledgements

This contribution was partially supported by the European Commission project SIMPATICO (H2020-EURO-6-2015, grant number 692819). We would like to thank the anonymous reviewers for their feedback.

References

Julian Brooke, Alexandra Uitdenbogerd, and Timothy Baldwin. 2016. Melbourne at semeval 2016 task 11: Classifying type-level word complexity using random forests with corpus and word list features. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 975–981, San Diego, California. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating Similar Languages: Evaluations and Explorations. In Proceedings of LREC.

David Kauchak. 2016. Pomona at semeval-2016 task 11: Predicting word complexity based on corpus frequency. In Proceedings of the 10th SemEval, pages 1047–1051.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Michal Konkol. 2016. Uwb at semeval-2016 task 11: Exploring features for complex word identification. In Proceedings of the 10th SemEval, pages 1038–1041.

Ludmila I Kuncheva, James C Bezdek, and Robert PW Duin. 2001. Decision Templates for Multiple Classifier Fusion: An Experimental Comparison. Pattern Recognition, 34(2):299–314.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and Human Baselines for Native Language Identification. In Proceedings of the BEA workshop.

Niloy Mukherjee, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. Ju_nlp at semeval-2016 task 11: Identifying complex words in a sentence. In Proceedings of the 10th SemEval, pages 986–990.

Gustavo Henrique Paetzold. 2016. Lexical Simplification for Non-Native English Speakers. Ph.D. thesis, University of Sheffield.

Gustavo Henrique Paetzold and Lucia Specia. 2016a. SemEval 2016 Task 11: Complex Word Identification. In Proceedings of SemEval.

Gustavo Henrique Paetzold and Lucia Specia. 2016b. SV000gg at SemEval-2016 Task 11: Heavy Gauge Complex Word Identification with System Voting. In Proceedings of the 10th SemEval, pages 969–974.

Ashish Palakurthi and Radhika Mamidi. 2016. liit at semeval-2016 task 11: Complex word identification using nearest centroid classification. In Proceedings of the 10th SemEval, pages 1017–1021.

Sarah E Petersen and Mari Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. In Proceedings of SLATE.

Robi Polikar. 2006. Ensemble Based Systems in Decision Making. Circuits and systems magazine, IEEE, 6(3):21–45.

Maury Quijada and Julie Medero. 2016. Hmc at semeval-2016 task 11: Identifying complex words using depth-limited decision trees. In Proceedings of the 10th SemEval, pages 1034–1037.

Luz Rello, Ricardo Baaz-Yates, Stefan Bott, and Horacio Sagüion. 2013. Simplify or Help?: Text Simplification Strategies for People with Dyslexia. In Proceedings of W4A.

Francesco Ronzano, Ahmed Abura’ed, Luis Espinosa Anke, and Horacio Sagüion. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In Proceedings of the 10th SemEval, pages 1011–1016.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In Proceedings of the ACL Student Research Workshop.

Lucia Specia. 2010. Translating from Complex to Simplified Sentences. Computational Processing of the Portuguese Language, pages 30–39.

Krzysztof Wróbel. 2016. Plujagh at semeval-2016 task 11: Simple system for complex word identification. In Proceedings of the 10th SemEval, pages 953–957.

Seid Muhie Yimam, Sanja Ġtajner, Martin Riedl, and Chris Biemann. 2017. Multilingual and Cross-Lingual Complex Word Identification. In Proceedings of RANLP.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. Macsar at semeval-2016 task 11: Zipfian and character features for complexword identification. In Proceedings of the 10th SemEval, pages 1001–1005.