The Whole is Greater than the Sum of its Parts: Towards the Effectiveness of Voting Ensemble Classifiers for Complex Word Identification

Nikhil Wani[†,⋆], **Sandeep Mathias**[⋆], **Jayashree Aanand Gajjam**[♣], **Pushpak Bhattacharyya**[⋆]
Center for Indian Language Technology
[⋆]Department of Computer Science and Engineering
[♣]Department of Humanities and Social Sciences
Indian Institute of Technology Bombay, India,
[†]nick.nikhilwani@gmail.com
[⋆]sam,pb{@cse.iitb.ac.in},[♣]jayashree_aanand@iitb.ac.in

Abstract

In this paper, we present an effective system using voting ensemble classifiers to detect contextually complex words for non-native English speakers. To make the final decision, we channel a set of eight calibrated classifiers based on lexical, size and vocabulary features and train our model with annotated datasets collected from a mixture of native and non-native speakers. Thereafter, we test our system on three datasets namely NEWS, WIKINEWS, and WIKIPEDIA and report competitive results with an F1-Score ranging between 0.777 to 0.855 for each of the datasets. Our system outperforms multiple other models and falls within 0.042 to 0.026 percent of the best-performing model's score in the shared task.

1 Introduction

Complex Word Identification (CWI) is an essential sub-task for Lexical Simplification. Lexical Simplification involves substituting a complicated word in the text with a more straightforward synonym. Figure 1 shows the pipeline for Lexical Simplification systems. It is geared for target population like non-native speakers, second-language learners, young learners, and people with language disabilities (like Aphasia and Alexia), with the aim of allowing them to comprehend the presented text completely.

The goal of the shared task is as follows: Given a target word (or phrase) and its context, we are to computationally determine if the target word is complex or not. Unlike the SemEval 2016 shared task, the target words here ***could have more than one word*** (e.g., *teenage girl*), and the context could stretch over multiple sentences.

The rest of the paper is organized as follows. In Section 2, we mention related work in the area of Complex Word Identification - in particular, the previous shared task at SemEval 2016

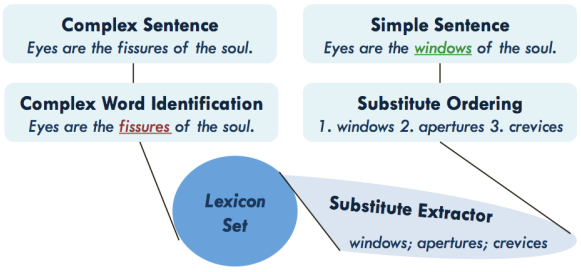

Figure 1: Lexical Simplification Pipeline

(Paetzold and Specia, 2016a). Section 3 describes the dataset of NLP BEA's CWI shared task at NAACL 2018. In Section 4, we describe our system, the features used, and our classification methodology. Moving along we then report our competitive results in Section 5 and discuss them in Section 6. We conclude by recapitulating our paper in Section 7 and identify future work that will be done.

2 Related Work

In SemEval 2016, 21 teams participated in a shared task on complex word identification (Paetzold and Specia, 2016a). The competition involved finding out whether a given word in a sentence was complex or not for a non-native speaker. The dataset used was completely in English.

In this task, the winning team used a soft voting-based approach from the outputs of 21 predictors (either classifiers, threshold-based, or lexical) (Paetzold and Specia, 2016b). This system was the best system according to the G-Score - an evaluation metric designed specifically for this task at SemEval 2016 (Paetzold and Specia, 2016a). The system with the best F1-Score made use of a threshold-based approach that marked a word as complex if its frequency in Simple Wikipedia is above a threshold (Wróbel, 2016).

Other systems at the SemEval 2016 shared

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 200–205
New Orleans, Louisiana, June 5, 2018.

Dataset	Total Sents.	Unique Sents.
NEWS-TRAIN	14002	1016
NEWS-TEST	2095	175
WIKINEWS-TRAIN	7746	652
WIKINEWS-TEST	1287	105
WIKIPEDIA-TRAIN	5551	387
WIKIPEDIA-TEST	870	61

Table 1: Description of the Dataset. The first column gives the dataset. The next column gives the total number of sentences. The last column gives the number of unique sentences.

task used SVM (Kuru, 2016; Choubey and Pateria, 2016; S P et al., 2016; Zampieri et al., 2016), Random Forest (Davoodi and Kosseim, 2016; Mukherjee et al., 2016; Zampieri et al., 2016; Brooke et al., 2016; Ronzano et al., 2016), Neural Networks (Bingel et al., 2016; Nat, 2016), Decision Trees (Quijada and Medero, 2016; Malmasi et al., 2016; Malmasi and Zampieri, 2016), Nearest Centroid classifier (Palakurthi and Mamidi, 2016), Naive Bayes (Mukherjee et al., 2016), threshold bagged classifiers (Kauchak, 2016) and Entropy classifiers (Konkol, 2016; Martínez Martínez and Tan, 2016).

The features used in most of the systems were common, such as length-based features (like target word length), presence in a corpus (like presence of the target word in Simple English Wikipedia), PoS features of the target word, position features (position of the target word in the sentence), *etc*. However, a few of the systems used some innovative features. One of them was the MRC Psycholinguistic database (Wilson, 1988) used by Davoodi and Kosseim (2016). Another system by Konkol (2016) used **a single feature** namely document frequency of the word in Wikipedia, for classifying using a maximum entropy classifier.

3 Datasets

For this shared task (Yimam et al., 2018), we used only the English monolingual dataset, which made use of data from a number of sources, such as News articles, WikiNews and Wikipedia articles. Table 3 shows details such as total sentences and the number of unique sentences that we computed across all the three datasets. The Wikipedia dataset consisted of sentences from Wikipedia articles. Likewise, the WIKINEWS dataset and the NEWS dataset contained sentences from news articles. However, the difference between the two is that the articles in the NEWS dataset were written by professional journalists, while lesser experienced writers wrote those in the WIKINEWS dataset.

In a majority of instances, the target words were just a single word. However, there were a few target words that were over a word long. Similarly, in most cases, the context was only one sentence, except for a few instances in which the context was as long as 3 - 4 sentences. The training datasets were annotated by 10 native and 10 non-native English speakers. Even if one amongst them found the word to be difficult, it was annotated as complex.

4 Methodology

In this section, we describe the experiment setup, such as the features used and provide analysis for their selection. This is followed by a detailed system overview which explains the system's architecture.

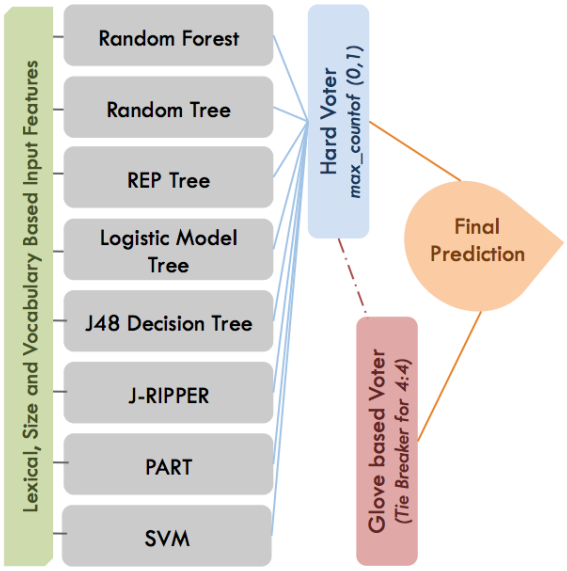

Figure 2: CWI System Architecture

4.1 Feature Sets

We investigated several *intuitive* properties of the target word such as its relevant lexical attributes, length properties and presence in certain word lists.

4.1.1 Lexical Features

The following features were extracted using WordNet (Fellbaum, 1998) for the target word:

- **Degree of Polysemy (DP):** Number of senses of the target word in WordNet (Fellbaum, 1998). This is operationalized by counting the number of Synsets of the target word in WordNet. Words with larger WordNet Synset sizes have several senses and were found to be more unclear.
- **Hyponym (Ho) and Hypernym (He) Tree Depth (TD):** These help in finding lexical relations. To find the position of the word in WordNet's hierarchical tree, we consider capturing its depth. General and simple words tend to be at the top of the tree. By computing the average depth among all the target-word Synsets, we count the number of Hyponyms and Hypernyms as a feature.
- **Holonym Count (HC) and Meronym Count (MC):** An alternative way to traverse Wordnet's hierarchical tree is by considering the relationship of the target word to its components (Meronyms) or to the things it is contained in (Holonyms). Holonyms tend to be more simple than meronyms because meronyms are usually more specific, compared to holonyms, as holonyms are a generalized word for a group of entities, while meronyms refer to specific entities in that group.
- **Verb Entailments (VE):** Verbs being action words often contain entailment relationships. For example, the act of roosting involves the act of sitting, so roosting entails sitting. Target words on average with multiple entailments were found to be relatively complex since they tend to be visually more vivid when trying to comprehend. Hence, the number of verb entailments of the target word was also part of our feature set.

4.1.2 Other Features

In addition to the lexical features, we also make use of size-based features and vocabulary-based features. These features are defined in Table 3.

4.2 System Overview

These input features are converged to the following eight calibrated classifiers, namely Random Forest, Random Tree, REP Tree, Logistic Model Tree, J48 Decision Tree, JRip Rules Tree, PART, and SVM, from a set of 16 classifiers (7 tree-based classifiers, 5 rule-based classifiers, 1 Bayesian classifier, 1 regression-based classifier, and 2 non-linear classifiers).

Classifier	Precision	Recall	F1-Score
Selected Classifiers			
Random Forest	0.792	0.781	0.787
J48 Decision Tree	0.777	0.777	0.777
Logistic Model Tree	0.778	0.762	0.770
REP Tree	0.768	0.765	0.766
Random Tree	0.796	0.717	0.754
SVM	0.745	0.780	0.762
PART	0.715	0.793	0.752
JRip Rules Tree	0.754	0.737	0.745
Rejected Classifiers *(F1 < 0.70)*			
Decision Table	0.739	0.652	0.693
Decision Stump	0.665	0.696	0.680
Hoeffding Tree	0.686	0.666	0.676
Logistic Regression	0.732	0.591	0.654
SMO	0.751	0.550	0.635
OneR	0.735	0.550	0.629
ZeroR	0.000	0.000	0.000

Table 2: Results of ten-fold cross-validation on the training for each of the classifiers on the ***complex class only***. This was used to choose our top classifiers.

SIZE-BASED FEATURES	
Feature	**Definition** *(Number of)*
Word Count (WC)	Words in the target word
Word Length (WL)	Letters in the target word
Vowels Count (VC)	Vowels in the target word
Syllable Count (SC)	Syllables in the target word
VOCABULARY-BASED FEATURES	
Feature	**Definition** *(Word is in)*
Ogden's Basic Lexicons (OB)	Ogden's Basic Word List
Ogden's Freq. Lexicons (OF)	Ogden's Frequent Word List
Barron's Lexicons (BW)	Barron's GRE Word List

Table 3: Size-based and Vocabulary-based features that we use.

These eight classifiers were chosen because they gave the best results on 10-fold cross-validation of the training set. We decided upon these classifiers since each of them had an F1-Score of the **complex class** in excess of 0.70. Table 2 describes the selected and rejected classifiers, along with their Precision, Recall and F1-Score on ten-fold cross-validation of the training data. Since the majority class was the non-

TEAM	DATASET		
	WIKINEWS	WIKIPEDIA	NEWS
camb	0.8430	0.8115	0.8792
ajason08	0.8368	0.7736	0.8625
nathansh	0.8329	0.7996	0.8706
nikhilwani	0.8213	0.7770	0.8554
dirkdh	0.8151	0.7816	0.8721
daalft	0.8050	0.7839	0.8391
TMU	0.7910	0.7621	0.8706
pom	0.7723	0.7460	0.8277
natgillin	0.7498	0.6690	0.8363

Table 4: F1-Score for each of the datasets for the top 10 teams on the corresponding test dataset. The highlighted row corresponds to our submission.

complex class, the ZeroR classifier has a Precision, Recall, and F1-Score of 0.

We use a hard voting approach to predict the class of the target word. If **more than 4 classifiers** classify the target word as either complex or simple, we assign the majority label to that word. In case of a 4-4 tie, (where 4 classifiers say the target word is complex and 4 say that it is simple), we use a word-embedding based classifier to act as a tie-breaker.

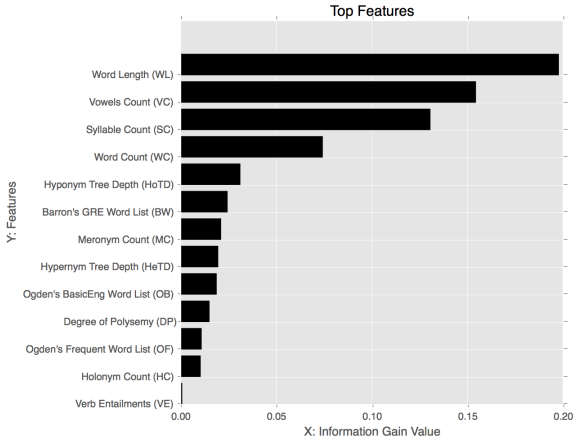

Figure 3: Feature significance observed by ranking them from highest to lowest using Attribute Evaluation based on Information Gain. The length of the bar corresponds to the actual Information Gain value.

For the word-embedding based classifier, we use the GloVe pre-trained word embeddings (Pennington et al., 2014). We first split the target into its constituent words (in most cases, it is a single word, but in a few cases, it is a phrase). We find the most similar word to each of the constituent words in the training set. If any of the given constituent words were tagged as complex, we target the target word as complex as well.

Out of 4252 test points to be classified, 173 times a tie occurred and the ensembled classifieres were unable to make a call. This is almost 4.06% of the predictions, which is significant in the larger scheme of things and further refines the hard voting.

5 Results and Analysis

In this section we discuss the results as well as reflect on the significance of each of the features for this task.

Table 4 gives the results of our experiments on the test set. From the results, our system is placed 4th in the WIKINEWS dataset, 5th in the WIKIPEDIA dataset, and 6th in the NEWS dataset.

Figure 3 delineates important features and ranks them according to their significance. Size based features namely - Word Length, Vowels Count, Syllable Count, Word Count were seen to constitute the first four topmost features. Another useful indicator of a complex word is its presence in Barron's GRE Word List, a list filled with the vocabulary level equivalent to a graduate college student.

6 Discussion

As it is evident from Tables 2 and 4, we see that individual classifiers do not work as well as ensembling them together, which agrees with the expression ***"The whole is greater than the sum of its parts"***. Classifier Ensembling would further prove to be an efficacy for contextual documents similarity-based binary classification tasks (Kanojia et al., 2017) which rely heavily on lexical features, as well as it should also potentially cross-pollinate to benefit probabilistic touch classification problems (Wani et al., 2017) where spatial and contextual information has been proven to be pivotal.

7 Conclusion and Future Work

In this paper, we describe our participation to NLP-BEA's CWI 2018 Shared Task at NAACL concerning Complex Word Identification. We presented and evaluated our system across three datasets and showed that Ensemble Classifiers with hard and GloVe Voting are effective by means of lexical, size and vocabulary features for identifying complex words.

As part of our future work, we plan to incorporate Parts of Speech (POS) tags, Named Entity Recognition (NER) tag and word position features to improve our existing effective system.

References

Joachim Bingel, Natalie Schluter, and Héctor Martínez Alonso. 2016. Coastalcph at semeval-2016 task 11: The importance of designing your neural networks right. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1028–1033, San Diego, California. Association for Computational Linguistics.

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.

Prafulla Choubey and Shubham Pateria. 2016. Garuda & bhasha at semeval-2016 task 11: Complex word identification using aggregated learning models. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1006–1010, San Diego, California. Association for Computational Linguistics.

Elnaz Davoodi and Leila Kosseim. 2016. Clac at semeval-2016 task 11: Exploring linguistic and psycho-linguistic features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 982–985, San Diego, California. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Diptesh Kanojia, Nikhil Wani, and Pushpak Bhattacharyya. 2017. Is your statement purposeless? predicting computer science graduation admission acceptance based on statement of purpose. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 141–145, Kolkata, India. NLP Association of India.

David Kauchak. 2016. Pomona at semeval-2016 task 11: Predicting word complexity based on corpus frequency. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1047–1051, San Diego, California. Association for Computational Linguistics.

Michal Konkol. 2016. Uwb at semeval-2016 task 11: Exploring features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1038–1041, San Diego, California. Association for Computational Linguistics.

Onur Kuru. 2016. Ai-ku at semeval-2016 task 11: Word embeddings and substring features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1042–1046, San Diego, California. Association for Computational Linguistics.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000, San Diego, California. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2016. Maza at semeval-2016 task 11: Detecting lexical complexity using a decision stump meta-classifier. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 991–995, San Diego, California. Association for Computational Linguistics.

José Manuel Martínez Martínez and Liling Tan. 2016. Usaar at semeval-2016 task 11: Complex word identification with sense entropy and sentence perplexity. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 958–962, San Diego, California. Association for Computational Linguistics.

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 International Workshop on Semantic Evaluation (SemEval-2016)*, pages 986–990, San Diego, California. Association for Computational Linguistics.

Gillin Nat. 2016. Sensible at semeval-2016 task 11: Neural nonsense mangled in ensemble mess. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 963–968, San Diego, California. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016a. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974, San Diego, California. Association for Computational Linguistics.

Ashish Palakurthi and Radhika Mamidi. 2016. Iiit at semeval-2016 task 11: Complex word identification using nearest centroid classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1017–1021, San Diego, California. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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 International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1034–1037, San Diego, California. Association for Computational Linguistics.

Francesco Ronzano, Ahmed Abura'ed, Luis Espinosa Anke, and Horacio Saggion. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016, San Diego, California. Association for Computational Linguistics.

Sanjay S P, Anand Kumar, and Soman K P. 2016. Amritacen at semeval-2016 task 11: Complex word identification using word embedding. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1022–1027, San Diego, California. Association for Computational Linguistics.

Nikhil Wani, Adarsh Patodi, and Sumit Singh Yadav. 2017. Probabilistic modeling of swarachakra keyboard for improved touch accuracy. In *Adjunct Proceedings of 16th IFIP TC.13 International Conference on Human Computer Interaction (INTERACT 2017 MUMBAI)*, pages 22–27, Industrial Design Centre, IIT Bombay, Mumbai, India.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

Krzysztof Wróbel. 2016. Plujagh at semeval-2016 task 11: Simple system for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 953–957, San Diego, California. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. Macsaar at semeval-2016 task 11: Zipfian and character features for complexword identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1001–1005, San Diego, California. Association for Computational Linguistics.