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, Indian Institute of Technology Bombay, India

nikhilwani@outlook.com, {sam,pb}@cse.iitb.ac.in, jayashree_aanand@iitb.ac.in

Overview

Goal: To determine if a given English target word or phrase in its context is complex or not for non-native English speakers.

Contribution:

- We use a set of eight classifiers based on WordNet lexical, size, and vocabulary features.
- Our system outperforms multiple other models and falls within 0.042 to 0.026 percent of the best model's score.

Motivation

1. Our proposed model will help people with language disabilities such as Aphasia and Alexia understand the text completely.
2. Complex Word Identification (CWI) is also an essential sub-task for Lexical Simplification.

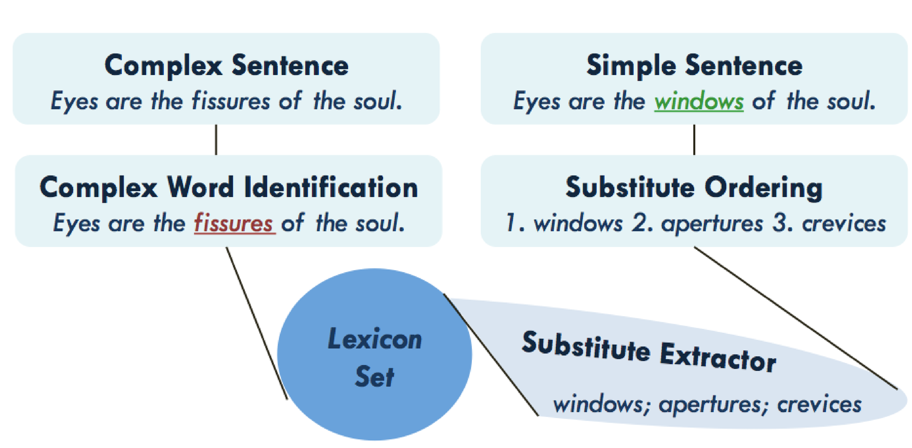

Figure 1: Lexical Simplification Pipeline

Shared Task Dataset

- 10 native and 10 non-native English speakers annotated a set of target words and phrases as complex or not, from Wikipedia, news reports, and amateur news reports (WikiNews)[3].

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: Shared Task Dataset - Descriptive Statistics

WordNet[1] Lexical Features

- **Degree of Polysemy (DP)** - Number of senses.
- **Hyponym (Ho) and Hypernym (He) Tree Depth (TD)** - Distance from the root (hypernym) and the longest path to a leaf (hyponym).
- **Holonym Count (HC) and Meronym Count (MC)** - Number of holoynyms and meronyms.
- **Verb Entailments (VE)** - Number of verb entailments.

Other Features

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 2: Size-based and Vocabulary-based features that we use.

System Architecture and Experiments

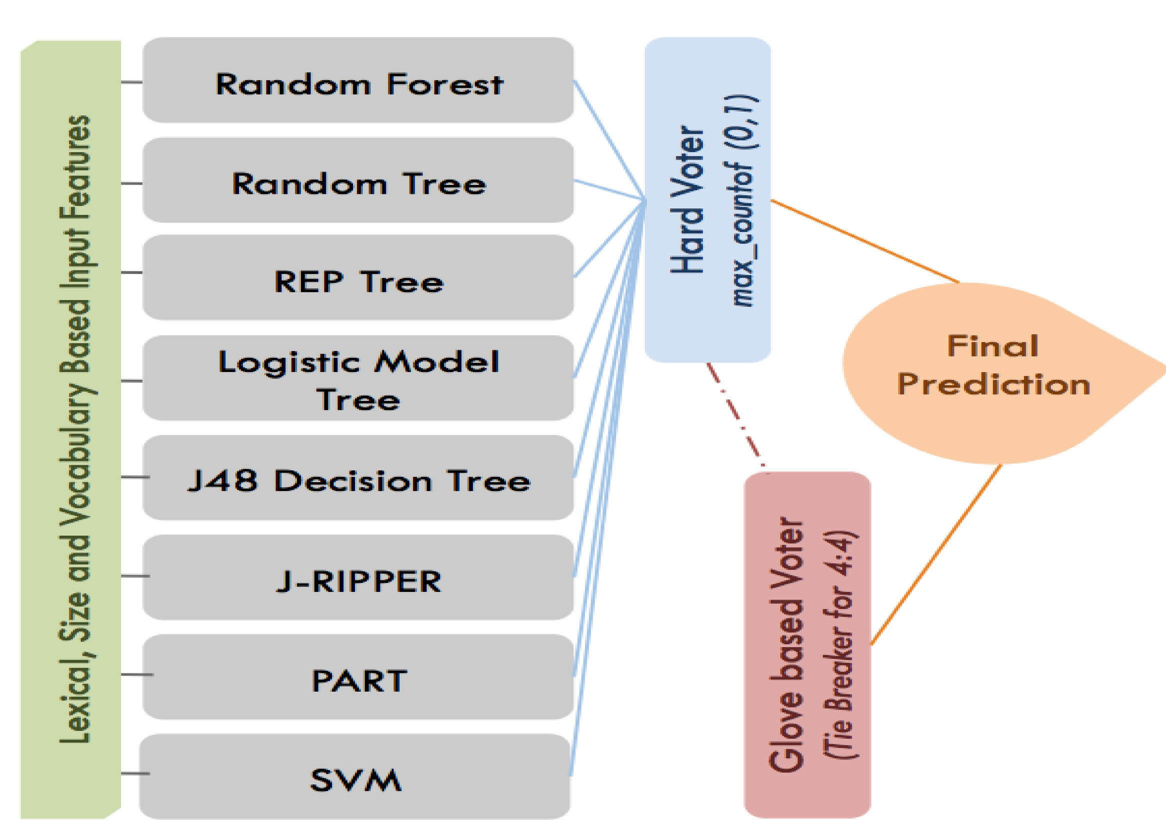

Figure 2: System architecture. Output of each of the classifiers goes to the voter. In case of a tie, we use GloVe [2]. Out of 4252 instance, a tie occurred 173 times.

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 3: 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.

References

[1] Christiane Fellbaum. *WordNet, Wiley Online Library*. Wiley Online Library, 1998.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. 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, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1162`.

[3] Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 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, June 2018. Association for Computational Linguistics.

Analysis and Discussion

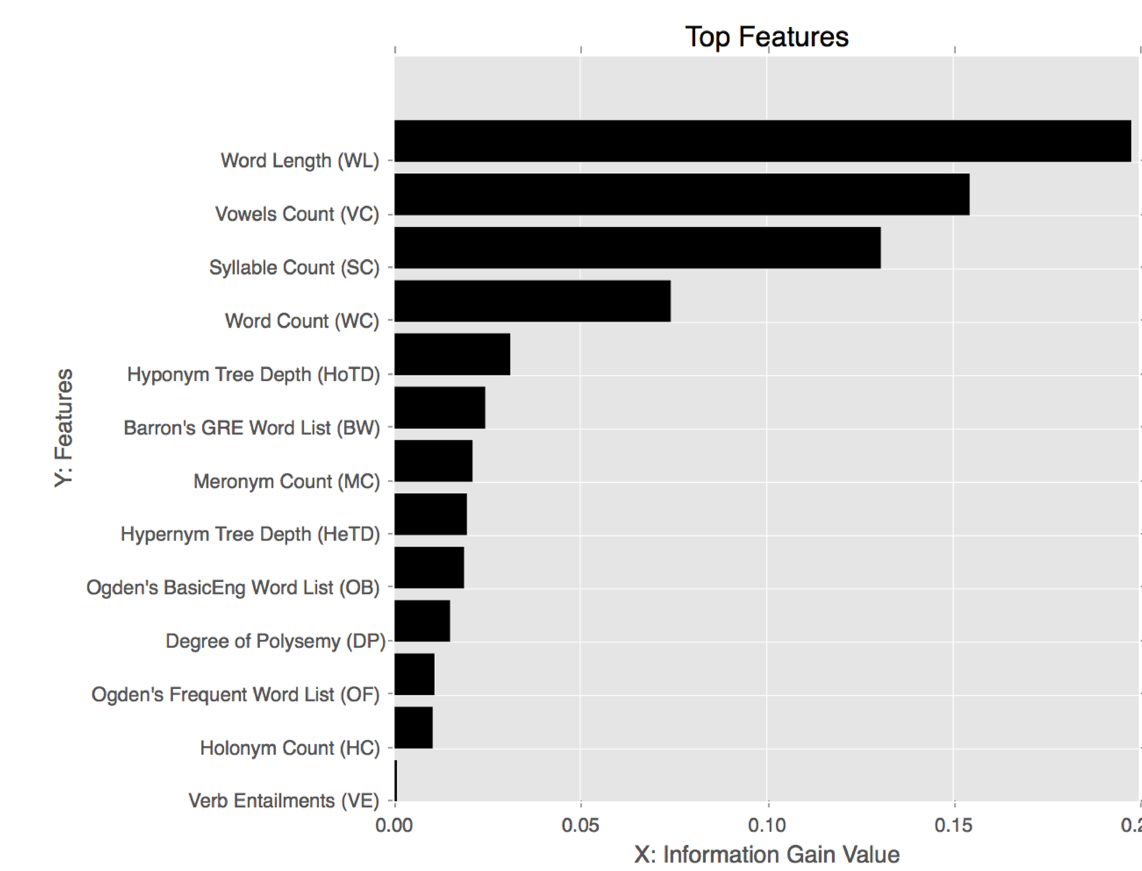

Figure 3: Feature significance observed by ranking them from highest to lowest using Attribute Evaluation based on Info-Gain

Results

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.

Conclusion and Future Work

- **Conclusion**: Ensemble classifiers with hard voting and GloVe is more effective than individual classifiers for CWI. Our Code is available here.[a]
- **Future Work**: Incorporation of Parts of Speech (POS) tags, Named Entity Recognition (NER) tags and word position features.

[a] `https://git.io/vh3G0`