JU_NLP at SemEval-2016 Task 11: Identifying Complex Words in a Sentence

Niloy Mukherjee, Braja Gopal Patra, Dipankar Das, and **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering, Jadavpur University
Kolkata, India
{niloymukherjee93,brajagopal.cse,dipankar.dipnil2005}@gmail.com
sivaji_cse_ju@yahoo.com

Abstract

The complex word identification task refers to the process of identifying difficult words in a sentence from the perspective of readers belonging to a specific target audience. This task has immense importance in the field of lexical simplification. Lexical simplification helps in improving the readability of texts consisting of challenging words. As a participant of the SemEval-2016: Task 11 shared task, we developed two systems using various lexical and semantic features to identify complex words, one using Naïve Bayes and another based on Random Forest Classifiers. The Naïve Bayes classifier based system achieves the maximum G-score of 76.7% after incorporating rule based post-processing techniques.

1 Introduction

Extensive research has been performed in the field of lexical simplification (Specia et al., 2012; Rello et al., 2013; Paetzold, 2015). Lexical simplification refers to identifying complex words and replacing them with lexically simple substitutes (Specia et al., 2012). The English lexical simplification task[1] was organized in the year 2012, in which the complex words were provided by the organizers.

The complex word (CW) identification is the first step towards the lexical simplification task. Understanding words which are not frequently used in any language is very difficult for non-native speakers. It may be challenging for a reader to interpret a particular word because it might be absent in his vocabulary. Also it may so happen that he knows the word but cannot comprehend it as he fails to capture the context it is used in. Generally, it is observed that the frequent use of complex words decreases the readability of the document. Thus, the complex word identification (CWI) task aims to classify those challenging words in a sentence with respect to a particular target audience.

For example, in the following sentence, the words in italics are complex words. These words are related to biology and are rarely used in our daily life. e.g. "The first *amniotes*, such as *Casineria*, resembled small lizards and evolved from *amphibian reptiliomorphs* about 340 million years ago."

Some research has been performed in CWI task in comparison to the lexical simplification (Shardlow, 2013; Paetzold, 2015). The important features which have been used previously in the CWI task are frequency thresholding and lexical matching etc. (Shardlow, 2013).

Some motivations of the CWI task are to understand the defining characteristics of the words which are challenging for non-native speakers to interpret. Another is assessing an individual's vocabulary limitations from the group he is a part of.

We have participated in the SemEval 2016-Task 11: Complex Word Identification[2] (Paetzold and Specia, 2016). The main goal of this task is to identify the complex words from English sentences. We identified highly correlated features and performed the classification using Naïve Bayes and Random Forest classifiers. After the classification, we used post-processing techniques with deterministic features to improve the F-Score of our system.

[1] https://www.cs.york.ac.uk/semeval-2012/task1.html

[2] http://alt.qcri.org/semeval2016/task11/

Proceedings of SemEval-2016, pages 911–915,
San Diego, California, June 16-17, 2016.

2 Dataset Description

Participants were provided with training and test datasets by the organizers of CWI task. A subset of words of a sentence are tagged as complex or noncomplex. The training and test datasets comprise of 2,237 and 88,221 instances respectively. The numbers of complex and noncomplex words are 706 and 1531 for the training dataset, whereas the number of complex and noncomplex words are 4,131 and 84,090 in the test dataset.

The training dataset was collected through a survey, in which 400 annotators were presented with 200 sentences. They were asked to select the words which they did not understand in terms of the meaning. Each of the words in the training dataset have been annotated by 20 distinct annotators. Even if one of the 20 annotators judged the word to be complex it has been tagged as complex. The test set has annotations made over 9,000 sentences by only one annotator (Paetzold and Specia, 2016).

3 Features

3.1 Data Pre-processing

The Stanford Parser[3] was used to get the lemma of the tagged words in the training dataset. Further, the lemmas of these words have been used to identify various features. The R (version 3.1.0)[4] software is used to collect various statistics and identifying the features which have high correlation with the complex or noncomplex class.

3.2 Part-of-Speech (POS)

We used POS tags of the words as a feature. The frequencies of corresponding POS tags of complex and noncomplex words are given in Table 1.

3.3 Hypernym and Hyponym

The main idea of the present approach is to find out the position of the words in the tree constructed by the WordNet.[5] Our hypothesis is that generic words being easier to understand are present at the top of the WordNet tree. Alternately specialized words which are difficult to understand are at the bottom of the tree. We used the number of hypernym and hyponym as features. For example, the word '*car*' has no hypernym and is tagged as noncomplex, whereas the word '*resemble*' has eight hypernyms and therefore is tagged as complex. We observed that 762 out of 1531 (around 50%) noncomplex words have no hypernyms in the training dataset.

POS	Complex	Noncomplex
NN	263	413
NNS	101	198
JJ	93	247
VBN	46	99
RB	45	103
VBD	40	101
Others	116	358

Table 1: POS tagging statistics

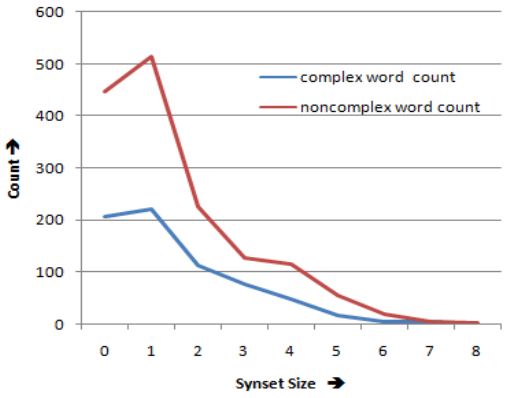

Figure 1: Statistics of synset size

3.4 Synset Size

The synset size is one of the important features which has been used to identify CWs in previous related work (Shardlow, 2013). We observed that the words with larger synset sizes have several senses and are generally ambiguous in nature. These words may be confusing for the readers and are considered complex. For example, approximately 73% of the words having synset size greater than equal to five were marked as complex in the training dataset. It can be observed from Figure 1 that the probability of a word being noncomplex is high when the synset size of that word is low.

[3] http://stanfordnlp.github.io/CoreNLP/
[4] https://www.r-project.org/
[5] https://rednoise.org/rita/reference/RiWordNet.html

3.5 Named Entity (NE)

Generally, NEs are understood by the non-native speakers. They are aware of currencies or nationalities, e.g. Dollars or British. We found 54 out of 78 (approximately 70%) NEs are noncomplex in the training data. We used the 7 class model of Stanford NER[6] to identify the NEs in the tagged words of the training and test dataset. We used NEs as a feature and for the post-processing as well.

3.6 Stopwords

We observed that determiners like a/an/the or conjunctions like or/and/but have a low probability of being complex. Thus, we used stopwords as a feature.

3.7 Syllable count

The words with a high number of syllables are difficult to pronounce and onerous to read too. The syllable count was calculated by the number of consonants present between contiguous chunks of vowels. We used syllable count as a feature to identify the CWs.

3.8 Most frequently used English words

We collected a list of most frequently used words in English language from the web.[7] Two lists were prepared, one containing top 2,000 words and the other containing top 5,000 words. We observed that the words present in the list of 'most frequently used words' have a lower chance of being complex.

3.9 Index of words

The index of each tagged word in a sentence is used as a feature.

Negative features: The length of the word was not used as a feature because a lot of noncomplex words in the training dataset were hyphenated and hence had higher length. The hyphenated words which are individually understandable should be considered as noncomplex.

[6]http://nlp.stanford.edu/software/CRF-NER.shtml

[7]http://functional-programming.it.jyu.fi/resources/word_list.txt

4 System Framework

The Naïve Bayes and Random Forest classifiers were implemented using the Weka tool.[8]

4.1 Evaluation

The performance of the systems was calculated using Accuracy, Precision, Recall, F-Score, and G-Score (Paetzold and Specia, 2016). The accuracy is calculated as:

Accuracy = (correctly classified instances) / (total instances).

The G-Score metric has been used to rank the systems. The G-score is the harmonic mean of Accuracy and Recall (Paetzold and Specia, 2016).

4.2 Post-Processing

Crawler: The specialized words of any particular subject are not generally understood by readers and are found to be complex. Thus, we prepared word lists of three specific topics (Biology, Geography, and Physics).

A web crawler was developed to collect specialized words from the glossaries of Biology,[9] Geography,[10] and Physics.[11] A total of 1327, 1689, and 273 number of words were collected for Biology, Geography and Physics, respectively. We observed that there are 48 CWs in the training dataset belong to the above glossaries. Thus, a word is tagged as complex, if it is found in any three of these glossaries.

Dictionary Module: We observed that the words not present in the English dictionary are tagged as complex. A python dictionary module *pyenchant*[12] (both US and UK) was used to identify the non-English words. If a word was not found in either of them, then it was tagged as complex.

Most frequently used English words: If a word is not found in the 5,000 word list, then it was tagged as complex.

Named Entity as noncomplex: The NEs which are identified as CWs by our system are re-annotated as noncomplex words. We also tagged positional words (such as 2nd/3rd/4th) as noncomplex.

[8]http://www.cs.waikato.ac.nz/ml/weka/

[9]http://www.phschool.com/science/biology_place/glossary/

[10]http://www.physicalgeography.net/glossary.html

[11]http://www.etutorphysics.com/glossary.html

[12]https://pypi.python.org/pypi/pyenchant

4.3 Results

We performed the 10-fold cross validation on the training dataset using the Random Forests classifier and achieved a F-Score of 0.53. Again, we applied the post-processing on the results obtained by the above system and observed an improvement of 0.04 in the F-Score. Thus, we presented all the results on the test dataset after implementing the post-processing techniques.

The Naïve Bayes and Random Forest classifiers based systems achieved the maximum accuracies of 0.767 and 0.795 respectively. However, the precision of both the systems are quite low (0.139 and 0.151). One of the main reasons is that the number of CWs in test dataset are quite low as compared to CWs in the training dataset (4.69% and 31.6% for test and training dataset, respectively). This happened because 20 annotators have annotated the training dataset and a word is tagged as complex if any one of the annotator annotated so. Whereas, a word in the test dataset is annotated by only one annotator.

The maximum recalls achieved by Naïve Bayes and Random Forest based systems are 0.767 and 0.73 respectively. Both the systems achieved almost similar recalls and accuracies because all the features used are biased towards complex words. The maximum F-Score achieved for the Naïve Bayes based system is 0.236 and that for Random Forest based system is 0.25. The detailed statistics of the system performances are given in Table 2. The confusion matrix for the above systems are given in Table 3.

The team **SV000gg** has achieved the first and second positions with G-scores of 0.774 and 0.773, respectively. The team **TALN** which came third with the maximum G-Score of 0.772 has used Random Forest classifier. They included the number of annotators who marked a particular word complex as a feature.

Our Naïve Bayes and Random Forest based systems achieved fourth and seventh position with the maximum G-scores of 0.767 and 0.761 respectively. There are two other systems namely **UWB** and **PLUJAGH** who have also achieved the fourth position. The team **UWB** uses Maximum Entropy classifiers and uses document frequencies of words in Wikipedia as the only feature. They obtain higher accuracy of 0.803, but have the same G-Score as our system. The team **PLUJAGH** achieved the same G-Score as our system, but they achieved the higher accuracy of 0.795. Their system learns the threshold of word frequencies in Wikipedia that maximizes the F-score over the joint dataset. Another system of the team **PLUJAGH** achieved the maximum F-Score of 0.353, but got quite low G-Score of 0.608 and obtained 22nd position.

	Acc.	Prec.	Rec.	F-Score	G-Score
NB	0.767	0.139	0.767	0.236	**0.767**
RF	0.795	0.151	0.730	0.250	0.761

Table 2: System performance (NB: Naïve Bayes, RF: Random Forest, Acc.: Accuracy, Prec.: Precision, Rec.: Recall)

		Predicted			
		NB		RF	
		0	**1**	**0**	**1**
Actual	**0**	64493	19597	67132	16958
	1	964	3167	1115	3016

Table 3: Confusion Matrix for systems

5 Conclusion and Future Work

In this paper, we have presented two systems for identifying the complex words in English. We believe that this problem will become increasingly important for lexical simplification. Our Naïve Bayes based system obtained the fourth position with the maximum G-score of 0.767.

In the training dataset, various stopwords within complex phrases were tagged as complex, because the annotator could not understand the context of that phrase. Thus, capturing complex phrase in a sentence is an interesting task and it would require context and n-gram level features. Apart from this, the feature set can be extended to build a model to identify the persons suffering from dyslexia.

Acknowledgments

The present work is supported by a grant from the project "CLIA System Phase II" funded by Department of Electronics and Information Technology (DeitY), Ministry of Communications and Information Technology (MCIT), Government of India. The second author is supported by "Visvesvaraya Ph.D.

Fellowship" funded by DeitY, Government of India.

References

Gustavo H. Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'16, San Diego, California.

Gustavo H. Paetzold. 2015. Reliable lexical simplification for non-native speakers. In *Proceedings of the NAACL-HLT 2015 Student Research Workshop (SRW)*, pages 9–16.

Luz Rello, Ricardo A. Baeza-Yates, and Horacio Saggion. 2013. The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2013)*, pages 501–512.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *ACL (Student Research Workshop)*, pages 103–109. Citeseer.

Lucia Specia, Sujay K. Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 347–355.