Complex Word Identification based on Word Features

Kenyonke
Department of Computer Science
University of Sheffield

Abstract

In this report, we present a features-based complex word identification(CWI) system by using SVM classifier. The system provides two monolingual language identified approaches, one is English and the other is Spanish. In real world, people may be confused by complex words with some specific features. Follow this rule, we try to extract some features of words/phrase which are possible to affect the complexity of the words/phrase. Moreover, the SVM classifier can be trained by extracted feature and label of words from the given training dataset.

1 Introduction

The task of CWI is automatically recognizing words or phrases which are not easy to be understood by target readers, such as non-native speaker, native speakers with low literacy levels and various types of reading impairments (**?**). Lexical complexity is a key role in reading comprehension for such readers. This work is the first step in lexical simplification system (**?**), which is aim to simplify sentence by substituting complex words for simple words and reorganizing the structure in order to improve text readability (**?**). Thus, increasing the accuracy of CWI plays a important role in lexical simplification system.

In real world, people could be confused by complex words(CWs) which lead to problems of reading. Also, CWs are difficult to define, because people has different abilities to understand words which means that words confuse someone may not make others confused. The Complex Word Identification (CWI) Shared Task 2018 provided English and Spanish datasets which define complex word by the number of native speakers and non-native speakers who marked the word as difficult.

The approach we used to build complex word classifier is creating a SVM classifier based on features of words/phrases. Given a training dataset, the system automatically extracts features such as frequency, length, syllable and synonym. After that, it maps words/phrases to vector space and train the binary with feature vectors and labels. Also, the classifier will be evaluated by macro-F1 score on experiments.

The report is organized as follows. Introductions for the project and its background knowledge are included in Section 1. Section 2 describes the analysis of baseline system. Section 3 details the implementation of our improved system. Finally, we give conclusions and inform future work.

2 Baseline system

The baseline system is composed of 3 functions: features extracting, training classifier and label prediction. At first, the system need to select which language is to be processed because English and Spanish are different in specific features. In features extracting, the baseline only look for the length of a word or a phrase, the word with more characters is more possible to be defined as complex word. Also, the binary classifier is based on logistic regression and user can change any classifier they want to improve the performance. The final step is label prediction, which is able to predict labels of words or phrases in testset by extracting their features to be input of the trained classifier.

3 Improved system

The improved model is composed of five step: language selection, extracting information from training set, training classifier, prediction and evaluation. The system provides complex words identification for monolingual English and monolingual Spanish. In baseline, we create a feature class for storing frequency of words and phrases in training set. Also, it provides a function to extract four

features of a word or phrase we need in training except the part of speech (POS). In processing of training, the training class finds out POS of word in sentence and loads others features from feature class to fit the SVM model. After that, the training class provide label of input words or phrases by inputting all the features of words or phrases. With the provided labels and gold-standard labels, the system automatically calculates scores of itself and print them.

In order to improve performance of the system, we tried to extract more features followed approaches provided by CWIG3G2() and Shardlow() to fit classifier. Features used by the system are shown as follow:

- N-grams Frequency: The frequency of n-gram in the corpus. In the system, we count from unigram frequency to 5-gram frequency in order to deal with phrases with multiple words. Also, frequency of words can be regarded as frequency of unigram. Words/phrases with higher frequency could be more familiar to people.
- Length: The number of characters of a word or a phrase. The word with more characters is more possible to be defined as complex word.
- Part Of Speech (POS): The POS influences the complexity of the word (**?**). Spacy library was used to find out POS of words in sentences.
- Synonyms Count: The number of synonyms a word has. It could affect the complexity of a word (**?**). For English the synonyms count is got from WordNet. For Spanish, we search information from online dictionary.
- Syllable Count: The number of syllables a word has. Words with more syllables could be complex (**?**). For English the Syllable count is got from the Carnegie Mellon Pronouncing Dictionary. For Spanish, we search information from online dictionary.

Another way to improve performance of the system is using alternative classifier. The improved system is based on Support Vector Machine (SVM), which has better performance in CWI than logistic regression.

In machine learning, SVM are supervised learning algorithms which is able to implement multiple classification by training (Wikipedia). Given a set of training samples with label "1"(complex) or "0"(simple), SVM training algorithm can build a model which predict labels "1" or "0" for new samples. Thus, the SVM we trained is a binary classifier. To implement such SVM, we use sklearn package, a free machine learning library. Also, RBF kernel and linear kernel were used respectively in SVM to improve accuracy.

4 Experiments

In experiments, the performance of CWI systems are mainly evaluated by overall macro-F1 score of both complex class and simple class. It is noted that the number of complex samples is not equal to the number of simple samples. So, using Macro-F1 as an overall performance metric is able to to avoid favouring.

The macro-F1 score of English datasets and Spanish datasets are presented in Table 1. On all the English dataset, our system reaches closed scores expect wikipedia test data. This is probably due to the imbalanced distribution of two classes of words in datasets. On Spanish datasets, the performance of Spanish complex words indentification is not as well as English words for the system.

Table 1: The result of improved system

testdata	macro-F1
English_Dev	0.79
English_Test	0.80
News_Dev	0.81
News_Test	0.81
Wikipedia_Dev	0.74
Wikipedia_Test	0.78
Wikinews_Dev	0.76
Wikinews_Test	0.77
Spanish_Dev	0.76
Spanish_Test	0.73

Accroding to the score in Table 2, the improved system works better than the baseline. The English CWI works as expected but Spanish CWI is not as expected. Even if more features are extracted and train the classifier, the performance of improved system are not improved much. For English words indentification, the result of improved system is efficiency increased.

We compare the perform between SVM and logistic regression by training them with a same training set in improved system. Table 3 shows

Table 2: Comparison between baseline and improved system

testdata	system	macro-F1
English	baseline	0.69
English	improved system	0.80
Spanish	baseline	0.72
Spanish	improved system	0.73

that the scorce of SVM is higher than logistic regression even if the performance of logistic classifier with more features is better than that with only one feature.

Table 3: Comparison between SVM and logistic regression

classifier	dataset	macro-F1
SVM	English	0.80
SVM	Spanish	0.75
logistic regression	English	0.77
logistic regression	Spanish	0.73

Also, We present leanring curves of the improved system based on SVM and LR in figure1 and figure2 respectively. Leanring curve is able to diagnosis bias and variance in order to reduce error. For SVM system, it is clear that the training score is still around the maximum value with small change. Also, with the increase of training examples the validation score is slowly raised and closed to training score. That means the model is trained to be more accurate, which is what we want. However, the LR system does not work as well as SVM and its validation seems the same in training.

Table 4: List of errors on prediction

word/phrase	language	correct label
gun	English	1
pact	English	1
dawn	English	1
flat	English	1
Oxley	English	1
aos	Spanish	1
clima	Spanish	1
sede	Spanish	1

For future work to improve the performance of

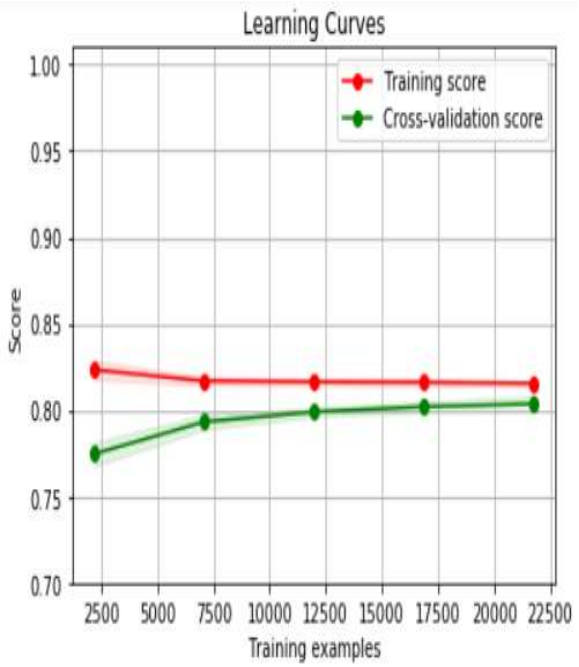

Figure 1: Learning Cures of SVM trained by English dataset

Figure 2: Learning Cures of LR trained by English dataset

improved system, we search some samples which the system can not prodict their correct labels (table 4). These words have short length in common, also, they are defined as complex word. However, they do not have frequency, syllables count, POS or synonyms count in common. Thus, there should be some features we did not use in the system. There are some ideas for future work to address them. The first is training model with gold-standard labels rather than binary labels, because many words in training set were deined as complex word which is considered as complex by only one annotators even they seems very simple. So, gold-standard label could be more accurate to define a complex word. The second solution is adding new features such as character sequence in words or phrases because the character sequence of complex words are more possible to be strange for readers.

Conclusions

In this report, we have introduced a feature based complex word indentification system. The idea of this system is extracting features such as n-grams frequency, length, part of speech, synonyms count and syllable count of words or phrases. Also, the SVM trained by extracted features from training set will able to predict whether a word or a phrase is complex for target population. The improved system have expected performance in English CWI, however the macro-F1 scorce on Spanish data is not as well as we expected. This result proves that English and Spanish have different features in defining complex words. Due to the limiation of knowledge on Spanish, it is hard for us to detail the diffenrence and get deeper insight to Spainsh processing.

In this class project, we got some ideas from the experiment which can improve the performance and fix some error in the system. In experiment, we found that tokenization of words by spacy package is not perfect which lead to some mistakes and affect the performance of system. Finding a better way to implement tokenization of datasets can improve performance of the system. Furthermore, as we had maped words and phrases into vectors by extracting features in training data, the system could be more powful by adding word embedding to existed features vector. Word2vec and RNN are helpful to implement such appraoch. The third idea is creating a more powful classifier. Beside logistic regression and SVM, there are many classifers in machine learning, such as naive bayes, decision tree and mulitiple classifer.

References