

Complex Word Identification using Neural Network

Jiaxin Zhang

University of Sheffield

JZhang124@sheffield.ac.uk

1 Introduction

This paper¹ describes the CWI (Complex Word Identification) Shared Task in 2018, which is about Complex Word Identification. This task aims to identify complex words in each monolingual language. 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 and Farzindar, 2010). It is important and necessary to identify complex words in the given texts. For non-native speaker, it would be easier for them to read the articles where complex words are identified and replaced by other much easier words. For native speakers with low literacy levels or various kinds of reading impairments, identifying complex words are also helpful to them.

Before the CWI Shared Task hold in 2018, there was a former task which is CWI shared task at SemEval 2016. There were 42 systems which have been submitted in the task. The highest G-Score was up to 77 % (Paetzold and Specia, 2016). In the previous task, Decision Trees and Ensemble methods have better results than Neural Network and embedding models.

This paper uses word embedding and linguistic features. The classification model is trained by neural network, which is a binary classification system. Finally, it got 0.83 for F1_Score in English and 0.75 for F1_Score in Spanish. This result is satisfying, and the flowing parts will elaborate the details.

¹The code could be find at <https://github.com/KnightZhang625/NLP-Code>

The paper introduces the training datasets in DataSets part firstly, then it talks about the baseline results. Feature choice is explained in Features part, and we talk about how to implement the training model in Experiment part. Finally, we compare the results of different combinations of features trained by either Logistical Regression or Neural Network.

Finally, we sum up the result and experience from our project, and talk something about future work.

2 DataSets

As this project only identifies complex words in English and Spanish, this part will only describe datasets for these two languages.

2.1 Description

English Dataset comprises news written by professionals, informal news from WikiNews and articles from Wikipedia. Spanish dataset is abstract from Spanish Wikipedia news. Each sentence in the datasets was read by both native and non-native speakers where they need to mark the words which would be difficult for children, non-native speakers or people with language disabilities to recognise.

There are 27,299 , 3,328, 4,252 for training, dev, test data respectively in English, and 13,750, 1,622, 2,233 training, dev, test data respectively in Spanish.

2.2 Example

The training data samples are showed as below: (Provided by CWI Shared Task 2018)

<ID> Both China and the Philippines
flexed their muscles on Wednesday. 31 51
flexed their muscles 10 10 3 2 1 0.25

<ID> Both China and the Philippines
flexed their muscles on Wednesday. 31 37
flexed 10 10 2 6 1 0.4

<ID> Both China and the Philippines
flexed their muscles on Wednesday. 44 51
muscles 10 10 0 0 0 0.0

There are 11 columns in each sentence from left to right. The second column shows the actual sentence. The fifth column represents the target word. The tenth and eleventh columns show the gold-standard label for the binary and probabilistic classification tasks. Since we are only interested in the binary classification task, the last column can be ignored (i.e. gold-standard probability). For the tenth column, 0 represents the target word as simple word, and 1 represents the target word as complex word. In the binary classification task, the aim is to predict the right label (0 or 1) for the test data.

3 Previous Work

There are two baseline systems, the second one is based on the first one.

The first simple classifier was given as a baseline. The classifier was Logistic Regression which was trained on two features from target words: length of characters and length of tokens. The baseline has achieved 0.69 and 0.72 F1_Score for English and Spanish. The scores for F1 were good, however they could be better. There is one possibility can be consider that the data may be too complicated to differentiate by a simple classification model. One way to deal with this is to construct features that are derived from polynomials of the original features. However the result were the same as the first.

As the traditional method may not get better results, the following work combine the word embedding and linguistic features together, and syllables and word senses were added into the original linguistic (i.e. Length) feature. The training method was changed to Neural Network where there were 1 hidden layer and 2 hidden layers for English and Spanish

respectively. The original Logistic Regression was also tested based on the new features, However, the F1_Scores are still a bit lower than the Neural Network.

4 Features

The combination of features consists of five different features: Word Embeddings, number of syllables, senses count, length of the word and tokens. Each feature will be explained individually as below:

4.1 Word Embedding

Each word has low dimension if words are represented only by linguistic features, so the feature vectors are sparse for this task, especially when the training model is Neural Network. So Mapping words to vectors with high dimensions becomes necessary. Word embedding maps the words to vectors, where each dimension represents different meaning. This is one reason why this project chooses word embeddings as words features. Another reason is that the features selected manually are insufficient , because we are not linguistic professionals. Although the training model will decrease the importance of the irrelevant features, it cannot increase influence of the non-added feature. In result, Word Embedding increased F1_Score tremendously. The implementation details will be explained in Experiments part.

4.2 Linguistic Features

Linguistic Features is some features which embody orthographic features of words. This project selected number of syllables, word senses count, length of the word and tokens as features. The number of syllables is the count of each words syllable. The word which has long syllables would have more chance to be a complex word. The count of word senses is the semantic senses that a word holds. If a word consists of many meanings or has various lemmas, this word would also have more probability to be a complex word. The final feature, length of the word and tokens which are original features in our baseline

model. It is very common that a complex word or phrase has a long length.

5 Experiment

5.1 Feature Extract

This project uses pre-trained word embeddings from Spacy. Spacy has pre-trained word embeddings for 7 languages (include English and Spanish). The embeddings for English are trained on OntoNotes, with GloVe vectors trained on Common Crawl, where each word vector contains 300 dimensions. And the embeddings for Spanish are trained on the AnCora and WikiNER corpus, where each word vector contains 50 dimensions.

The feature of the count of the syllables is obtained from the Pyhyphen package. And the feature of the senses of the word is from NLTK WordNet.

The sizes of the vector for each word in English and Spanish are 304 dimensions and 54 dimensions respectively.

5.2 Training

In this part, This project has tested different combinations of features in both logistical regression and neural network.

The feature combinations are: 1.the length of words and tokens; 2.All the linguistic features; 3.Word Embeddings + Linguistic Features.

For Neural Network training model, the loss function is cross entropy loss. The model use Adam to optimise the parameters, where the learning rate is 0.001, and the hyperparameters for β_1 and β_2 are 0.99 and 0.9 respectively. There are 1 hidden layer and 2 hidden layers for English and Spanish respectively. The iteration times are 10,000 and 15,000 for English and Spanish respectively.

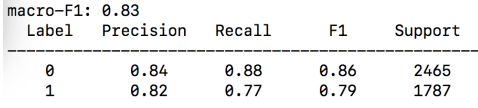
6 Result

The figures below illustrate the F1_Score for English and Spanish:

We let F1, F2, F3 refer to Length of words and tokens features, all linguistic features, Word Embeddings + Linguistic Features respectively. M1 refers to Logistic Regression, and M2 refers to Neural Network.

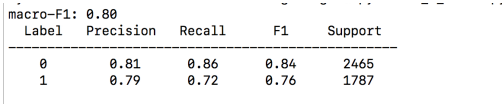
English:

The graphs below are the screenshot for Neural Network and Logistic Regression.



macro-F1: 0.83				
Label	Precision	Recall	F1	Support
0	0.84	0.88	0.86	2465
1	0.82	0.77	0.79	1787

Figure 1: Neural Network using Word Embeddings and Linguistic Features



macro-F1: 0.80				
Label	Precision	Recall	F1	Support
0	0.81	0.86	0.84	2465
1	0.79	0.72	0.76	1787

Figure 2: Logistic Regression using Word Embeddings and Linguistic Features

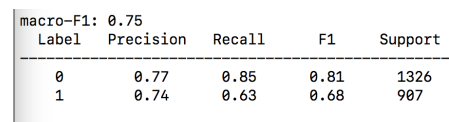
The table below shows the F1_Score for each combination in two training models. We can find that Word Embeddings increases the score tremendously, and the Neural Network model performs slightly better than the Logistical Regression model. However there is a trade-off, because the time taking by the Neural Network is much longer than the Logistic Regression. If we prefer speed to accuracy, it is reasonable for us to use Logistic Regression.

	F1	F2	F3
M1	0.69	0.72	0.80
M2	0.69	0.72	0.83

Table 1: F1_Score for English

Spanish:

The graphs below are the screenshot for Neural Network and Logistic Regression.



macro-F1: 0.75				
Label	Precision	Recall	F1	Support
0	0.77	0.85	0.81	1326
1	0.74	0.63	0.68	907

Figure 3: Neural Network using Word Embeddings and Linguistic Features

macro-F1: 0.72				
Label	Precision	Recall	F1	Support
0	0.74	0.89	0.81	1326
1	0.77	0.54	0.63	907

Figure 4: Logistic Regression using Word Embeddings and Linguistic Features

The table below shows the F1_Score for each combination in two training models. However, F1_Score for Spanish is lower than it for English, even they use the same features and training model. The size of training data for Spanish is half smaller than that for English may be a reason. And the word embedding dimensions for Spanish is 50 compared with 300 dimensions for English, which may result in losing some features. But the combination of Word Embeddings and Linguistic features still performs the best in Neural Network than others.

	F1	F2	F3
M1	0.71	0.71	0.72
M2	0.70	0.72	0.75

Table 2: F1_Score for Spanish

Learning Curves:

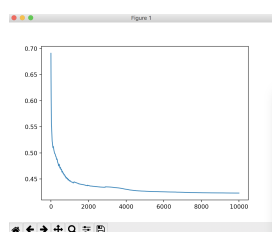


Figure 5: Learning Curve For English after 10,000 iterations

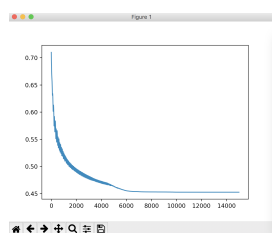


Figure 6: Learning Curve for Spanish after 15,000 iterations

7 Conclusion

This paper has described what we have done for the CWI(Complex Word Identification) Shared Task in 2018. In the project, different combinations of the features and different training model have been tested. Finally, we draw a conclusion that Neural Network performs better than the logistic regression and Word Embeddings feature plays the crucial role. Our model has achieved higher F1_Score than the baseline model, especially in the English task which gets 0.83 F1_Score.

There are some improvements which we can do in the future. As we have used the word embeddings, an important feature for the word embedding is that we can calculate the similarity between the words easily. We can add the similarity between the words as one feature, which may improve the performance. We use the pre-trained word embeddings, one reason is that training a new word embeddings takes too much time, another is that the size of training data is too small, especially for Spanish. There is one shortcoming that this pre-trained word embeddings may not capture the specific features of words in our specific training data. Once we could obtain much larger training data, we could learning word embeddings by our own. This may capture some certain features of complex words in our training data.

References

- Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. pages 560–569.
- Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: A corpus analysis.
- Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with hter. pages 33–41.