Class project: Complex Word Identification

Jagoda Karpowicz
Department of Computer Science
University of Sheffield, UK
jakarpowicz1@shef.ac.uk

Abstract

This paper contains documented developed implementation and findings of binary classification task for the monolingual English and monolingual Spanish in the second edition of **Complex Word Identification Task 2018**.

1 Introduction

1.1 Description

Dataset **CWIG3G2** from (Yimam et al., 2017) is in a form of sentences with annotations for words classified as difficult to understand for children, people with language disabilities and non-native speakers by annotators. The Spanish dataset is taken from Wikipedia pages while English dataset consists of both Wikipedia pages and professional and non-professional written news taken from WikiNews. Each sentence was annotated by native and non-native speakers. English one contained 27299 annotations in training dataset and 3328 in development dataset while in Spanish we had 13750 annotations in training dataset and 1622 in development dataset. The aim of this task is to extract language features *(lexical, semantic and others)* and predict if word is simple or not using binary classificator. Accuracy is measured by macro-averaged F1.

1.2 Task importance

Complex Word Identification (CWI)(Yimam et al., 2018) is the task considered as a prime step in lexical simplification (LS) process aim at identifying which words are considered complex by a given target population (Shardlow, 2013). An effective CWI approach should keep to the minimum number of simple word replacement that leads to semantic or gramatical errors whilst pointing out which our can be changed on simplified synonims possible to understand by people with lower language abilities. As proved in past papers (Paetzold and Specia, 2013(Paetzold and Specia, 2016)) ignoring this step have severe impact on a quality of the simplification.

2 Baseline system describtion:

Baseline system primarily uses basic features like normalized length of the target words and number of words in target. Implemented algorithm consisted of logistic regression with default parameters. During development process additional features were added.

2.1 Features

Features partially were inspired by papers from previous edition (Francesco Ronzano et al., 2016) (Ronzano et al., 2016) (Paetzold and Specia,2016a) (Paetzold and Specia, 2016).

2.1.1 Basic features

These features are based on position and length of target words.

- **Length of the target words** This featured were normalized by declared average length of word in both languages.
- **Number of words** Some of annotations referes to phrases - not words.
- **Position of the target word** Feature indicating position of the target word in the whole sentence divided by number of words in target phrase

2.1.2 Morphological features

These features covers morphology of the words like typical characters or n-grams of characters. All that features were learned in training process.

- **Trigram of characters** Feature representing every trigram of characters in a target words (without whitespace) for every target word longer than 2. Feature counts were normalized by length of the word minus two.

- **Bigram of vowels** Feature represents every possible bigram of vowels with their counts divided by length of the word.
- **Preffixes** Feature learned by iterating through predefined list of common prefixes in both languages searching for the longest match. Feature normalized by number of words in a target.
- **Suffiixes** As above, feature learned on predefined list of common suffixes.
- **Acents (Spanish)** This feature refers only to Spanish since words can contain accents (only one possible). Feature was divided by number of words in a target.

2.1.3 Semantic features

The features below trying to capture semantics of the word and possible plural meanings. WordNet (Princeton University, 2010) implementation was used.

- **Number of possible part of speech tags (English)** Number of possible part-of-speech -tags (nouns, verbs, adjective etc.) divided by number of words in a target.
- **Number of possible meaning (English)** Number of posible meaning (synsets) for each posible part-of-speech tag divided by number of words in a target. Feature representing conceptual-semantic and lexical relations.

2.1.4 Lexical features

These feature are learned from external corpora.

- **Simple-word corpora** Based on the Dale-Chall Word List (Dale and Chall, 1948) containing 3000 simple words known in reading by at least 80 percent of the children in Grade 5.
- **Word frequency based on corpora** Based on word frequency data for English implemented in **wordfreq** (Speer et al., 2017) that combines data from different sources (also Wikipedia). The bigger lists cover words that appear at least once per 100 million words.

2.2 Model

As a baseline only logistic regression model was used. Logistic regression classifies an observation into one of two classes.

3 Improved system motivation and description

3.1 Models

Since a baseline have not achieved satisfying performance in a second step system was expected to increase accuracy of prediction via de choosing another model with parameters tuning. The following models (from scikit-learn) for classification problem were checked (with pointed out the best configuration of parameters found):

- **Logistic Regression** (verbose 0.01)
- **AdaBoost**
- **Support Vector Machines** (linear kernel)
- **K-nearest neighbors** 15 neighbors, brute force search, euclidean distance)
- **Multi-layer Perceptron** (hidden layer sizes: 100, 150, 50, activation of tanh, adaptive learning rate)
- **Decision Tree** (depth of 5, entropy)
- **Random Forest** (depth of 13, entropy)

3.2 Findings

- **Decision Tree** with maximum depth of 5 and criterium of entropy was the best choice for Spanish.

- **Random Forest** with maximum depth of 13 and criterium of entropy was the best choice for English.

4 Results

4.1 English

4.1.1 Accuracy of predictions

Table 1: Accuracy

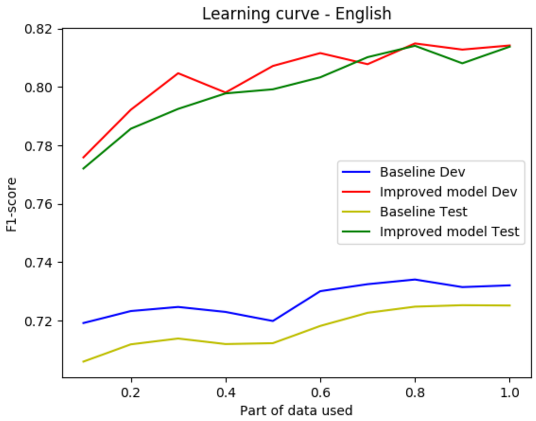
Accuracy (F1 measure)		Development data		Test data	
%	**Size of data**	**Baseline**	**Improved model**	**Baseline**	**Improved model**
100%	27299	0.7321	0.8142	0.7252	0.8138
90%	24569	0.7315	0.8128	0.7253	0.8081
80%	21839	0.7341	0.8149	0.7248	0.8141
70%	19109	0.7325	0.8078	0.7227	0.8102
60%	16379	0.7301	0.8116	0.7182	0.8033
50%	13649	0.7199	0.8072	0.7123	0.7992
40%	10919	0.7230	0.7981	0.7120	0.7978
30%	8189	0.7247	0.8047	0.7139	0.7925
20%	5459	0.7233	0.7922	0.7119	0.7857
10%	2729	0.7192	0.7759	0.7060	0.7721

The best accuracy for English was 81.5%

4.1.2 Explanation of improvements

None of the additionally feature were added during improvement. What helped in prediction was using a Decision Tree instead of Logistic Regression that search for a single linear decision boundary in your feature space, whereas a Decision Tree creates have a non-linear one.

4.1.3 Accuracy with different sample size



As we can see, improvements significantly increased accuracy visible in both development an testing dataset. Outcome is more prone to the data size for improved model, nonetheless number of records affected also accuracy of baseline. For both developing an testing dataset the improvement achieved is about 8%.

4.1.4 Examples of wrong predictions

Below prediction of target word where first numbers refers to label predicted in baseline system, second number to improved system and last one the golden label.

Predicted wrongly as easy:

- **('0', '0', '1', 'trial')** - most probably cause model learned that it is easy from corpora,
- **('0', '0', '1', 'mean')** - most probably because it is short word,
- **('0', '0', '1', 'madness')** - most probably model learned that suffix 'ness' is not difficult,
- **('0', '0', '1', 'Kelvin')** - most probably cause it is short and morphologically easy while it is a part of name,

Predicted wrongly as a difficult:

- **('1', '1', '0', 'support bases')** - most probably cause of length,
- **('1', '1', '0', 'Sheikh Abbas al-Laham')** - long word while it is name of a person,
- **('1', '1', '0', 'Petersburg')** - long word while it is name of a city,
- **('1', '1', '0', 'attention')** - most probably model learned suffix 'tion' as difficult,

4.2 Spanish

4.2.1 Accuracy of predictions

Table 2: Accuracy for Spanish

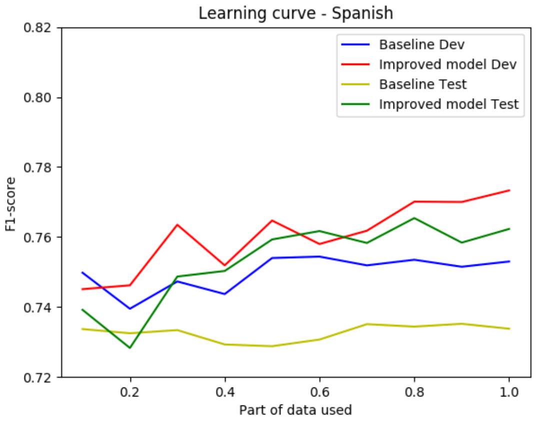
Accuracy (F1 measure)		Development data		Test data	
%	**Size**	**Baseline**	**Improved model**	**Baseline**	**Improved model**
100%	13750	0.7530	0.7733	0.7338	0.7623
90%	12375	0.7515	0.7700	0.7352	0.7584
80%	11000	0.7535	0.7701	0.7344	0.7654
70%	9625	0.7519	0.7618	0.7351	0.7583
60%	8250	0.7544	0.7580	0.7307	0.7617
50%	6875	0.7540	0.7647	0.7288	0.7593
40%	5500	0.7437	0.7519	0.7293	0.7503
30%	4125	0.7473	0.7635	0.7334	0.7487
20%	2750	0.7395	0.7462	0.7325	0.7283
10%	1375	0.7498	0.7451	0.7337	0.7392

The best accuracy for Spanish was 77%

4.2.2 Explanation of improvements

Prediction improvement was achieved with Random Forest classifier instead of Logistic Regression due to advantage over second algorithm that class labels roughly lie in hyper-rectangular regions not linear one. In that case, Random Forests turned out to be the better keeping balance between precision and overfitting than simple Decision Tree.

4.2.3 Accuracy with different sample size



As we can see, improvements increased accuracy more in testing dataset. Differences depending on a sample size are less smooth and the tendency of increasing accuracy while enlarging data is not always the case. Nonetheless, still it can be seen that generally the better outcome is obtain with more data. Difference between accuracy for baseline and improved model is smaller - about 3%.

4.2.4 Examples of wrong predictions

Predicted wrongly as easy:

- **('0', '0', '1', 'notables')** most probable cause it has not captured any morphological and semantic features,
- **('0', '0', '1', 'ritual')** most probably because target was short,
- **('0', '0', '1', 'laser')** -most probably because corpora suggested that it is easy word,
- **('0', '0', '1', 'albergaba')** -model has not captured that it is past tense.

Predicted wrongly as a difficult:

- **('1', '1', '0', 'aplicaciones')** most probable cause target word is long,
- **('1', '1', '0', 'corresponda')** most probably because target is long and have accent,
- **('1', '1', '0', 'conquistaba')**-most probably cause is long and have bigrams of vowels beside,
- **('1', '1', '0', 'Minneapolis')**-most probably because NER was not used to enable model to learn easy names

5 Conclusion

From the **CWI 2018 task** we have learnt that by choosing reasonable and adjusted features, external lexicons, sufficient amount of data and more advanced algorithms with appropriate parameters we can improve accuracy of prediction. Obtained result, especially for English, are good enough to assume that model is able to predict difficult words correctly. The features that contributed the most to the better outcome was those based on external lexicons and morphological ones. **Decision Tree** and **Random Forest** turned out to be the most powerful algorithms. Nonetheless, there is still a big room for improvements. In my opinion **Name Entity Recognition** would help the most, as well as **Word Embbeding**, using textbfbigger and better corpora, taking into account **psycholinguistic features** and prediction of target words separately would still increase accuracy.

6 References

References

E. Dale and J. S. Chall. 1948. A formula for predicting readability.

Gustavo Paetzold and Lucia Specia. 2016. Sv000gg at semeval-2016 task 11:heavy gauge complex word identification with system voting.

Princeton University. 2010. "about wordnet." wordnet.

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.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words.

Robert Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2017. Luminosoinsight/wordfreq: v1.7.

Seid Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Stajner, Anais Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018.

Seid Yimam, Sanja Stajner, Martin Riedl, and Chris Biemann. 2017. Multilingual and cross-lingual complex word identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017,pages 813?822, Varna, Bulgaria. INCOMA Ltd.*