MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for Complex Word Identification

Marcos Zampieri[1,2], Liling Tan[1], Josef van Genabith[1,2]
[1]Saarland University, Germany
[2]German Research Center for Artificial Intelligence, Germany
{first.last}@uni-saarland.de

Abstract

This paper presents the *MacSaar* system developed to identify complex words in English texts. *MacSaar* participated in the SemEval 2016 task 11: Complex Word Identification submitting two runs. The system is based on the assumption that complex words are likely to be less frequent and on average longer than words considered to be simple. We report results of 82.5% accuracy and 27% F-Score using a Random Forest Classifier. The best *MacSaar* submission was ranked 8th in terms of F-Measure among 45 entries.

1 Introduction

Complex Word Identification (CWI) is the task of automatically identifying complex words in texts. It is considered a sub-task carried out in most lexical simplification pipelines (Paetzold and Specia, 2015). In this step, complex words, which are likely to be difficult words for readers and language learners, are identified so they can be substituted for simpler ones (Specia et al., 2012; Shardlow, 2013). Lexical simplification methods are usually integrated into text simplification systems developed for a particular target population (e.g. people with reading impairment or dyslexia, language learners, etc.) (Siddharthan, 2014).

Given a sentence, a CWI system is trained to identify words which are considered by readers to be complex. To give an example, let us consider the following sentence extracted from the SemEval CWI task training set:

(1) Leo took an <u>oath</u> of <u>purgation</u> concerning the <u>charges</u> brought against him , and his opponents were <u>exiled</u>.

Taking Example 1 into account, the task of the CWI system is to assign as complex the four underlined words, namely: *oath, purgation, charges*, and *exiled*. But what makes these words complex and not, for example, *opponents* or *took*?

In the lexical simplification literature, the term *complex* is a synonym for difficult or complicated. For practical purposes, we consider as complex words those that were assigned by a pool of human annotators, provided by the organizers of the CWI task, as difficult to be understood due to several factors that we will discuss in this paper. This is a readability notion that is not necessarily related to intrinsic linguistic phenomena (e.g. word formation).[1]

1.1 Motivation

MacSaar participated in the CWI SemEval task interested in two aspects of complex words. The first one is related to communication principles, or in other words, what makes words complex or simple to readers. One of our assumptions is that complex words tend to be less frequent in general language corpora than simple words. The second aspect is language learning. Lexical and text simplification methods are very important to produce simpler texts

[1]It is important to note that the definition of *complex* used here is different from that used in Morphology, where complex words are defined as compound words or words composed of multiple morphs as opposed to *simplex* words which are words with no affixes and not part of compounds (e.g. *happy* is a simplex word and *unhappiness* is a complex one) (Adams, 2001).

Proceedings of SemEval-2016, pages 926–930,
San Diego, California, June 16-17, 2016. ©2016 Association for Computational Linguistics

targeted at language learners which facilitate reading comprehension.

In communication theory, the *cooperative principle* states that interlocutors cooperate and mutually accept one another to be understood in a particular way to optimize each interaction (Grice, 1975). Interactions should take what Grice describes as the four maxims into account: *quality, quantity, relevance*, and *manner*.

We relate the *maxim of manner* to the usage of simple words that a learner will hear more frequently than complex words. Therefore, it should be possible to determine whether a word is complex or simple by observing Zipfian frequency distributions computed from suitable text corpora (Zipf, 1949). Another aspect to consider is the length of the words. Words that are more frequent tend to be shorter as noted by Zipf: '*the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences*' (Zipf, 1935). That said, our approach takes both frequency and word length into account to determine whether a word is complex or not.

Finally, another aspect that we take into account is the difficulty in vocabulary acquisition that is related to the spelling of complex words (Xu et al., 2011; Dahlmeier et al., 2013). Educational applications that are tailored towards non-native speakers use character-level n-grams to identify possible spelling errors that language learner make. Thus making character combinations another interesting aspect to be consider in this task.

2 Related Work

CWI is a sub-task included in many lexical and text simplification systems. Lexical simplification, as the name suggests, focuses only on the substitution of complex words for simpler words in texts whereas text simplification comprises also the modification of syntactic structures to improve readability. Most text simplification systems also contain a lexical simplification module or component which often relies on the accurate identification of complex words for subsequent substitution. The three tasks are therefore inseparable.

Both lexical and text simplification approaches have been widely investigated. They have been applied to different languages, examples include: Basque (Aranzabe et al., 2012), Italian (Barlacchi and Tonelli, 2013), Portuguese (Aluísio et al., 2008), Spanish (Bott et al., 2012), and the SemEval lexical simplification task for English (Specia et al., 2012).

To the best of our knowledge, very few methods have focused solely on complex word identification prior to the CWI shared task. An exception is the work by Shardlow (2013) which compared different techniques to identify complex words.

3 Methods

3.1 Task and Data

The SemEval 2016 Task 11, Complex Word Identification (CWI) is a binary text classification task at the word level. Systems are trained to attribute a label of either 1 (for complex words) or 0 (for simple words) to each word in a given sentence. There are no borderline cases or gradation, all words are either complex of simple.

A tokenized data set containing English sentences annotated with the complex or simple label for each word was provided. The training set contained 2,237 sentences, and the test set contained 88,221 sentences. The shared task website[2] states that: 'the data was collected through a survey, in which 400 annotators were presented with several sentences and asked to select which words they did not understand their meaning'. There was no information of whether annotators were English native speakers.

The proportion of training vs. test instances makes the task more challenging than other similar shared tasks which provide much more training than test instances (Tetreault et al., 2013; Zampieri et al., 2015), a common practice in text classification tasks.[3]

3.2 Approach

Given the motivation described in Section 1.1, we approach the CWI task using word frequency and character-level n-gram features.[4] To emulate a language learner exposure to English, we use newspa-

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

[3] The Chinese grammatical error diagnosis (CGED) shared task (Yu et al., 2014) is an exception. See the discussion in Zampieri and Tan (2014).

[4] Our implementation is open source and it can be found on: https://github.com/alvations/MacSaar-CWI

pers text from the English subsection of the DSL Corpus Collection (Tan et al., 2014) to compute the word frequencies and n-gram probability used to train our classifier.

The features used are explained in detail in the following sections and summarized in Table 1.

Type	Feature
Zipfian	Zipfian Frequncy (ZipfFreq)
	True Frequncy (TrueFreq)
Orthographic	Word Length (no. of chars)
Difficulty	Word-level Trigrams Density
	Sentence Length (no. of words)
	Sentence-level Trigram Density

Table 1: Features used in *MacSaar*

3.2.1 Zipfian Features

We model the language learners perspicuity by using insights from Zipfian properties of human language. Zipf (1949) predicts that the frequency of an element from a population of n elements, *ZipfFreq*, is defined as follows:

$$ZipfFreq(word) = \frac{1}{k^s H_{n,s}} = \frac{1}{k_{word}} \quad (1)$$

where k is the rank of the word sorted by most frequent first, s is the exponent characterizing the distribution, n is the vocabulary and size $H_{n,s}$ is the generalized harmonic number i.e. the sum of the reciprocals of the size of vocabulary. In the simplest case, where we assume that the harmonic number and exponent to be 1, we compute *ZipfFreq* by taking the inverse of the the rank of a word.[5]

The Zipfian frequency is a hypothetical estimate of the nature of word frequency in natural language. To account for the true frequency of the word, we calculate the non-smoothed probabilities of the count of a word divided by the number of tokens in the corpus. Formally:

$$TrueFreq(word) = \frac{count(word)}{N} \quad (2)$$

where N is the number of (non-unique) words in the corpus.

[5]In the actual implementation of our submission, we have taken the percentile of the word rank, i.e. the product of the rank of the word and the inverse of number of words in the vocabulary, $|n|$. Empirically, they have the same effect in a classification since $|n|$ is a constant.

3.2.2 Character-based Features

To measure orthographic difficulty, we model word complexity by computing its (i) word length and (ii) sum probability of the character trigrams (normalized by the sum of all possible trigrams within the word). Intuitively, we could skip the normalization of the n-grams since we can assume that longer words are more complex. But we have the word length feature to account for the length of words, so the normalization of the n-grams probabilities would account for density of the n-gram probabilities independent of the length of the word.

Additionally, we computed (iii) sentence length and (iv) sum probability of the character trigrams of the sentence to account for contextual orthographic complexity with respect to the word-level spelling complexity. These sentence-level features are similar to those used in Native Language Identification (Gebre et al., 2013; Malmasi and Dras, 2015; Malmasi et al., 2015b).

As a meta-feature that captures both word and sentential level spelling complexity, we use the proportion of word to sentence orthographic difficulty by taking the ratio of the aforementioned features (ii) and (iv).

3.2.3 Classifiers

We trained 3 different classifiers using the features described in Table 1: a (i) Random Forest Classifier (RFC), (ii) Nearest Neighbor Classifier[6] (NNC) and (iii) Support Vector Machine[7] (SVM).

Nearest neighbor classifiers usually work well when the distribution between the training set data points are dense and similar to (or representative) of test set. Since there is a limitation of two official submissions, we only submitted the output generated by RFC and SVM.[8]

4 Results

The shared task organizers reported 45 submissions to the CWI task (including baseline systems). An overview of the task containing the complete scores

[6]RFC and NNC trained using Graphlab Create https://dato.com/products/create/ with default parameters (without tuning)

[7]SVM trained using Scikit-Learn (Pedregosa et al., 2011)

[8]SVM has been shown to perform well for large text classification tasks (Malmasi and Dras, 2014; Malmasi et al., 2015a).

Rank	Team	System	Accuracy	Precision	Recall	F-score	G-score
1	PLUJAGH	SEWDFF	**0.922**	0.289	0.453	**0.353**	0.608
2	LTG	System2	0.889	0.220	0.541	0.312	0.672
3	LTG	System1	0.933	**0.300**	0.321	0.310	0.478
4	MAZA	B	0.912	0.243	0.420	0.308	0.575
5	HMC	DecisionTree25	0.846	0.189	0.698	0.298	0.765
6	TALN	RandomForest_SIM	0.847	0.186	0.673	0.292	0.750
7	HMC	RegressionTree05	0.838	0.182	0.705	0.290	0.766
8	**MACSAAR**	RFC	0.825	0.168	0.694	0.270	0.754
9	TALN	RandomForest_WEI	0.812	0.164	0.736	0.268	0.772
10	UWB	All	0.803	0.157	0.734	0.258	0.767
11	PLUJAGH	SEWDF	0.795	0.152	0.741	0.252	0.767
12	JUNLP	RandomForest	0.795	0.151	0.730	0.250	0.761
13	SV000gg	Soft	0.779	0.147	**0.769**	0.246	**0.774**
14	**MACSAAR**	SVM	0.804	0.146	0.660	0.240	0.725
15	JUNLP	NaiveBayes	0.767	0.139	0.767	0.236	0.767

Table 2: The top 15 out of 45 systems in the shared task, ranked by their F-score.

obtained by all participants is available in the shared task report (Paetzold and Specia, 2016).

In Table 2 we include the top 15 submissions ranked by F-Score. We report results in terms of Accuracy, Precision, Recall, F-score, and G-score. The best scores for each metric are presented in bold.[9] Our best performing system (RFC) achieved 82.5% accuracy and 27% F-Score. Our second system (SVM), scored 2.1 percentage points accuracy and 3.0 percentage points less than the one using RFC.[10] Our best submission was ranked 8th in the CWI task in terms of both F-Score and G-Score.

We observed that some systems were trained to obtain good Recall and G-Score, for example the system ranked 13th by team SV000gg, while others obtained high Accuracy, for example the systems by teams LTG (System1), PLUJAGH, and MAZA which obtained accuracy scores higher than 90%. No system delivered a balanced combination of both scores which confirms the difficulty of this task.

Finally, as to the performance of the NNC system, we tested the NNC model on the gold data and this system achieved 75.9% accuracy and 11% F-score. As expected, it did not perform well because of the split between training and test set.

[9] G-score is the harmonic mean between Accuracy & Recall.

[10] In the official CWI task scores (Paetzold and Specia, 2016), our second system is referred to as NNC even though it used an SVM. This occurred because we substituted the output of the NNC for the SVM but were unable to change the entry's name.

5 Conclusion and Future Work

The two *MacSaar* submissions were ranked on the top half of the table, among the top 15 out of 45 entries, in the SemEval-2016 Task 11: Complex Word Identification (CWI). Our best system using a Random Forest Classifier was ranked 8th in terms of both F-score and G-Score. This indicates that the performance we obtained can be comparable to other state-of-the-art systems for this task.

More than a good performance, we showed that the use of Zipfian features are a good source of information for this task. The frequency of occurrence and word length in complex and simple words are two interesting variables to be investigated in future work. By looking at the relationship between word frequencies and word length Piantadosi et al. (2011) states that word lengths are optimized for efficient communication and that 'information content is a considerably more important predictor of word length than frequency'. In our approach we did not take information content into account and we would like to investigate this in the future.

Another interesting, and to a certain extent surprising, outcome is that the SVM classifier did not outperform RFC using the same set of features. Due to its architecture, SVM is well-known for performing well in binary classification tasks and we would like to look analyse the most informative features and to perform error analysis to investigate the reasons for SVM's poor performance in this task.

Acknowledgments

We would like to thank Gustavo Paetzold and Lucia Specia for organizing the CWI shared task. We also thank the anonymous reviewers and Shervin Malmasi for their feedback.

Liling Tan is supported by the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Valerie Adams. 2001. *Complex words in English*. Routledge.

Sandra M Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of DocEng*.

Marıa Jesús Aranzabe, Arantza Dıaz de Ilarraza, and Itziar Gonzalez-Dios. 2012. First approach to automatic text simplification in basque. In *Proceedings of the NLP4ITA Workshop*.

Gianni Barlacchi and Sara Tonelli. 2013. Ernesta: A sentence simplification tool for childrens stories in Italian. In *Proceedings of CICLing*.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of Coling*.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the BEA Workshop*.

Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving native language identification with tf-idf weighting. In *Proceedings of the BEA Workshop*.

Paul Grice. 1975. Logic and conversation. In *Syntax and Semantics*, pages 41–58. Academic Press, New York.

Shervin Malmasi and Mark Dras. 2014. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of EMNLP*.

Shervin Malmasi and Mark Dras. 2015. Multilingual Native Language Identification. In *Natural Language Engineering*.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015a. Arabic Dialect Identification using a Parallel Multidialectal Corpus. In *PACLING*.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015b. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the BEA Workshop*.

Gustavo Henrique Paetzold and Lucia Specia. 2015. LEXenstein: A framework for lexical simplification. In *Proceedings of ACL*.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of SemEval*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *Proceedings of the ACL Student Research Workshop*.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of SemEval*.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of BUCC*.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the BEA Workshop*.

Wei Xu, Joel Tetreault, Martin Chodorow, Ralph Grishman, and Le Zhao. 2011. Exploiting syntactic and distributional information for spelling correction with web-scale n-gram models. In *Proceedings of EMNLP*.

Liang-Chih Yu, Lung-Hao Lee, and Li-Ping Chang. 2014. Overview of grammatical error diagnosis for learning chinese as a foreign language. In *Proceedings of ICCE*.

Marcos Zampieri and Liling Tan. 2014. Grammatical error detection with limited training data: The case of chinese. In *Proceedings of ICCE*.

Marcos Zampieri, Liling Tan, Nikola Ljubešic, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the LT4VarDial Workshop*.

George Kingsley Zipf. 1935. *The Psychobiology of Language*. Houghton Mifflin.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort.* Addison-wesley Press.