USAAR at SemEval-2016 Task 11: Complex Word Identification with Sense Entropy and Sentence Perplexity

José Manuel Martínez Martínez and Liling Tan
Universität des Saarlandes
Campus, Saarbrücken, Germany
j.martinez@mx.uni-saarland.de
liling.tan@uni-saarland.de

Abstract

This paper describes an information-theoretic approach to complex word identification using a classifier based on an entropy based measure based on word senses and sentence-level perplexity features. We describe the motivation behind these features based on information density and demonstrate that they perform modestly well in the complex word identification task in SemEval-2016. We also discuss the possible improvements that can be made to future work by exploring the subjectivity of word complexity and more robust evaluation metrics for the complex word identification task.

1 Introduction

Complex Word Identification (CWI) is the task of automatically identifying difficult words in a sentence.

> It is an important subtask *prior to* the textual/lexical simplification task that *pertains to* the *substitution* of *abstruse* words with *lucid* variants which can be *apprehended* by a wider *gamut* of readers (Siddharthan, 2006; Specia et al., 2012; Shardlow, 2013).

The aim of the CWI task is to annotate the difficult words as shown in the underlined examples in the previous paragraph, such that a lexical simplification system can produce the following sentence:

> It is an important subtask *before* the textual/lexical simplification task that *concerns* the *replacement* of *difficult* words with *simpler* variants which can be *understood* by a *wider* range of readers.

Lexical simplification is a specific case of lexical substitution where the complex words in a sentence are replaced with simpler words.

Historically, lexical substitution was conceived as a means to examine the issue of the appropriateness of a fixed word sense inventory in the word sense disambiguation task the "sense" of a polysemous word is correctly identified given a context sentence (Kilgarriff, 1997; Palmer, 2000; Hanks, 2000; Ide and Wilks, 2007; McCarthy and Navigli, 2009). By allowing fluidity in the "sense" inventory and by quantifying how much the systems were able to generate good substitutes, these lexical substitutes would have built a word sense cluster of words that may not be covered by a set of pre-defined words in a sense inventory, e.g. Princeton WordNet (Miller, 1995) and Open Multilingual WordNet (Bond and Paik, 2012).

2 Entropy and Perplexity

Entropy is an information-theoretical measure of the degree of indeterminacy of a random variable[1]. In simpler words, entropy measures how unpredictable an event is likely to occur (Shannon, 1951).

For the case of complex word , we can also assume that the degree of word ambiguity contributes to its level of unpredictability which determines its complexity. We define the degree of word ambiguity as the number of possible senses a word can have,

[1] https://www.encyclopediaofmath.org/index.php/Entropy

Proceedings of SemEval-2016, pages 958–962,
San Diego, California, June 16-17, 2016.

more specially the number of synsets of a lemma of the target word as recorded in the Princeton WordNet. Formally, we define the ***sense entropy*** of a word, $H(word)$, as such:

$$H(word) = -\sum_{k=1}^{n} p(sense_k) \log p(sense_k) \quad (1)$$

where n is the number of possible sense of a word and $p(sense_k)$ is the probability of sense given the context sentence where the word occurs. We assume a uniform distribution across all senses of a word, thus we assign $1/n$ to the $p(sense_k)$ variable.

Perplexity is inverse measure of entropy that measures how predictable an event is likely to occur. Intuitively, if a complex word appears in a sentence, the sentence would become less common and less predictable, yielding a higher sentence perplexity score. Mathematically, we define the ***sentence perplexity***, $2^{H(sentence)}$ as follows:

$$2^{H(sentence)} = 2^{-\sum_{i=1}^{N} p(word_i) \log p(word_i)} \quad (2)$$

where N is the number of words in the sentence and $p(word_i)$ is the unigram probability of the word generated from a modified Kneser-Ney language model (Chen and Goodman, 1999).

3 Experimental Setup

The dataset for the CWI task in SemEval-2016 is annotated at word level with binary labels; 1 for complex and 0 for non-complex.

	Train	**Test**
# Sentences	200	8,929
# Labels	2,237	88,221
# Lemma	1,903	20,016
# Synsets	1,617	12,989
% WN Cover	84.97	64.89
% Complex	31.86	4.68

Table 1: CWI Task Dataset for SemEval-2016.

Table 1 presents the corpus statistics of the dataset provided for the CWI Task. The organizers have decided to emulate the limited human language capacity with a small training set and a large testing set that reflects the relatively larger proportion of text that a human will encounter in reality. However, we do note the stark difference between the percentage of complex words in the training and test data; it skews towards words being annotated with the non-complex labels.

To compute the sense entropy, we annotated the dataset with lemmas using the PyWSD lemmatizer (Tan, 2014) and reference the lemmas to the Princeton WordNet. The training and testing set comprise 2,237 and 88,221 words respectively. Of the annotated words, the training and testing set has 1,903 and 20,016 unique lemmas and the WordNet covers 84.97% and 64.89% of these lemmas respectively. When a lemma is not covered by WordNet, we assign an entropy of 0 that indicates that the lemma's complexity is easily predictable and the classifier would assign the majority label to the word.

To compute the sentence perplexity as presented in the previous section, we use the English Wikipedia section of the SeedLing corpus (Emerson et al., 2014) and the news articles from the DSL Corpus Collection (Tan et al., 2014) to train the language model using the KenLM tool (Heafield et al., 2013). On average, there are 11 annotated words per sentence and every word in the same sentence shares the same sentence perplexity.

Using both the sense entropy and sentence perplexity as features, we train a boosted tree binary classifier (Friedman, 2002) using the Graphlab Create[2] machine learning toolkit to identify the word complexity.

Interestingly, when we use the raw number of senses instead of sense entropy as a feature on various machine learning classifiers, the number of senses were uninformative and the classifiers either labels all words as complex or all words as non-complex.

4 Results

We submitted 2 systems to the CWI task in SemEval-2016 (Paetzold and Specia, 2016), one using only the sense entropy (`sentropy`) and another that includes the sentence perplexity feature (`entroplexity`).

The complex word classification would be evaluated based on classic (i) `accuracy`, (ii) `precision`, (iii) `recall`, (iv) `F-score`. In ad-

[2]https://dato.com/products/create/

dition to the harmonic F-score between the precision and recall, the organizers reported the harmonic mean between the accuracy and recall, dubbed `G-score`[3].

Since the accuracy score computes the percentage of true positive labels globally, it might be more indicative to read the accuracy scores given the highly skewed dataset (<5% of the test set is labelled as complex).

Table 2 presents the comparative results between our systems, 4 systems that ranked top in F-score and G-score and 2 baseline systems that uses threshold frequencies that best separate complex from simple words learned from the English and Simple English Wikipedias.

PLUJAGH-SEWDFF uses frequency thresholding techniques, they consider any word that occurs less than 147 times in the simple English Wikipedia to be complex. LTG-System2 uses a decision tree classifier trained using similar threshold features. SV000gg uses a soft and hard voting ensemble to combine 23 different systems that includes threshold-based and lexicon-based techniques and machine learning classifiers based on 69 distinct morphological, lexical, semantic, collocational and nominal features.

Compared to the top systems, our system has performed modestly and our `Sentropy` system outperforms the thresholding baselines. We note that our accuracy and precision scores are relatively competitive as compared to the top systems but our recall is distinctly lower which affects the F- and G-scores. Possibly, we could improve the system by using a word sense disambiguation tool that provide the sense probabilities instead of assuming uniform probabilities across all senses, especially when word senses are often dictated by the most common sense of the word given the context sentence.

Intuitively, we can expect the `Entroplexity` system with sentence-level perplexity to underperform in this particular test set because the variance of the perplexity measures are low since all words within the same sentence attain the same sentence perplexity. For a dataset where there are more training sentences, the feature could perform better.

[3]This should not be confused with the G-Measure that is used to measure cluster similarities (Fowlkes and Mallows, 1983).

5 Discussion

5.1 Subjectivity of Word Complexity

From the example in the introduction, we see the subjectivity of word complexity and how it may vary from speaker to speaker. Arguably *substitution* and *replacement* could have been of equal word complexity depending on the speaker's level of English proficiency. Although the word *variant* could easily be considered simple for a native French/German speaker learning English where the equivalence *Variante* exists in his/her native language, it might have been considered a complex word for other second English language speaker.

The annotations from the CWI task training set were collected from 20 annotators over a set of 200 sentences. A word is labelled complex if any one of the annotator deems it to be complex, while the testing set was annotated by 1 annotator.

To explore the reader-based subjectivity in word complexity identification, we suggest that future work on CWI explores reader-specific annotations and models user-specific annotations. In this respect, readers' meta-data such as their native and non-native languages, country of residence, etc. could potentially be more telling in predicting their English proficiency and identifying complex word catered to specific readers or groups of readers.

5.2 Evaluation Metrics

Complex Word Identification is a novel task and possibly the standard F-score and accuracy measures might not be reflective of the task difficulty or the system efficiency. Given the binary nature of the classification task, we suggest the use of Matthews correlation coefficient (Matthews, 1975) that measures the measures the correlation coefficient between the observed and predicted binary labels, which can be viewed as a variant of the chi-square coefficient[4].

It measures the discordant relations between the true and false positives and negatives and avoids the need to optimize the systems based on either accuracy or precision but a healthy fusion of both. The coefficient value ranges from -1 to +1 where +1, 0

[4]http://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html

Teams-System	Accuracy	Precision	Recall	F-Score	G-Score
PLUJAGH-SEWDFF	**0.922**	**0.289**	0.453	**0.353**	0.608
LTG-System2	0.889	0.220	0.541	0.312	0.672
SV000gg-Soft	0.779	0.147	0.769	0.246	**0.774**
SV000gg-Hard	0.761	0.138	**0.787**	0.235	0.773
USAAR-Sentropy	0.869	0.148	0.376	0.212	0.525
Baseline-Wiki Threshold	0.536	0.084	0.901	0.154	0.672
Baseline-SimpleWiki Threshold	0.513	0.081	0.902	0.148	0.654
USAAR-Entroplexity	0.834	0.097	0.305	0.147	0.447

Table 2: Comparative Results between our Systems, the Top Systems and Threshold-based Baselines in SemEval-2016 CWI Task.

and -1 respectively represents perfect, random and inverse predictions.

6 Related Work

Although the lexical simplification/substitution task is well-studied, the complex word identification task has mostly been discussed as an anterior subtask.

Devlin and Tait (1998) implemented a lexical substitution system that ***considers all words as complex words*** and generated the simpler variant of the words by referencing the most frequent synonym of the word from the WordNet synsets and SUBTLEX corpus (Brysbaert and New, 2009).

Another method to identify complex words is to use the Zipfian nature of language by ***thresholding frequencies and classify words that occurs below a certain threshold as complex***. Zeng et al. (2005) and Elhadad (2006) applied the thresholding method to the medical domain to identify technical terms that non-experts would find it difficult to read, the complex terms and varying frequencies correlate with the word difficulty scores elicited from questionnaires (Zeng-Treitler et al., 2008). Similarly, (Zampieri and Tan, 2014) used n-gram based frequency threshold to identify complex words that has caused second language Chinese learners to make errors in their essays.

Other than the heuristics described above, previous studies had also used **supervised machine learning algorithms** and data annotated with binary labels for each words in the training corpus. Malmasi et al. (2016) use Zipfian word ranks and character n-grams features to train a random forest, an SVM and a nearest neighbour classifier to predict word complexity.

Shardlow (2013) compared various techniques to identify complex words, viz. (i) treating every word as complex, (ii) thresholding frequency using the mean of the thresholds discovered through the highest accuracies achieved across cross-validations folds of the training and (iii) an SVM classifier using word-level and character-level (orthographic and phonemic) frequencies and the number of synsets of each words.

While the '*everything is complex*' technique achieved the highest recall, the SVM classifier scored the best precision[5]. The coefficients in his SVM classifier presented the sense feature as the weakest while the frequency features indicated higher correlations with the binary label distribution[6]. In comparison, our sense entropy system is based solely on the number of senses per word reported modest results in the CWI task.

7 Conclusion

In this paper, we presented our systems submitted to the complex word identification task in SemEval-2016. We introduced the notion of sense entropy that measures the unpredictability of a word based on its number of senses and used it as a feature to identify complex word . The implementation of our system is released as an open source tool available on `https://github.com/alvations/entroplexity`.

[5]The SVM outputs were significantly different from the other two techniques.

[6]Although coefficient values indicates a feature 'strength' anecdotally and the coefficients does not necessarily explain the true effect on the label decision because each coefficient is influenced by other variables.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

References

Francis Bond and Kyonghee Paik. 2012. A survey of wordnets and their licenses. In *GWC 2012 6th International Global Wordnet Conference*, volume 8, page 64.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–393.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*, pages 161–173.

Noémie Elhadad. 2006. Comprehending technical texts: predicting and defining unfamiliar terms. In *AMIA Annual Symposium, pp. 239-243. Washington, DC.*

Guy Emerson, Liling Tan, Susanne Fertmann, Alexis Palmer, and Michaela Regneri. 2014. Seedling: Building and using a seed corpus for the human language project. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 77–85, Baltimore, Maryland, USA.

Edward B Fowlkes and Colin L Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, 78(383):553–569.

Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1):205–215.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Nancy Ide and Yorick Wilks. 2007. Making sense about sense. In *Word sense disambiguation*, pages 47–73. Springer.

Adam Kilgarriff. 1997. I dont believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Shervin Malmasi, Marcos Zampieri, and Liling Tan. 2016. Macsaar: Zipfian and character-level features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Diana McCarthy and Roberto Navigli. 2009. The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Gustavo Henrique Paetzold and Lucia Specia. 2016. Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*.

Martha Palmer. 2000. Consistent criteria for sense distinctions. *Computers and the Humanities*, 34(1-2):217–222.

Claude E Shannon. 1951. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109.

Lucia Specia, Sujay Kumar 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.

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 The 7th Workshop on Building and Using Comparable Corpora*.

Marcos Zampieri and Liling Tan. 2014. Grammatical error detection with limited training data: The case of chinese. In *Proceedings of the 1st Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA'14), Nara, Japan*.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. In *Biological and Medical Data Analysis*, pages 184–192. Springer.

Qing Zeng-Treitler, Sergey Goryachev, Tony Tse, Alla Keselman, and Aziz Boxwala. 2008. Estimating consumer familiarity with health terminology: a context-based approach. *Journal of the American Medical Informatics Association*, 15(3):349–356.