ISSN: 0974-6471, Vol. 10, No. (3) 2017, Pg. 612-617

Oriental Journal of Computer Science and Technology

Journal Website: www.computerscijournal.org

An Automated Complex Word Identification from Text: A Survey

JASPREET SINGH[1], GURVINDER SINGH[2] and RAJINDER SINGH VIRK[3]

[1] Research Scholar, Department of Computer Science, Guru Nanak Dev University Amritsar. Punjab, India.
[2]Dean Faculty of Engineering, Department of Computer Science, Guru Nanak Dev University Amritsar. Punjab, India.
[3]Prof. and Head, Department of Computer Science, Guru Nanak Dev University Amritsar. Punjab, India.

Abstract

Complex Word Identification (CWI) is the process of locating difficult words from a given sentence. The aim of automated CWI system is to make non-native English user understand the meaning of target word in the sentence. CWI systems assist second language learners and dyslexic users through simplification of text. This study introduces the CWI process and investigates the performance of twenty systems submitted in the SemEval -2016 for CWI. The G-score measure which is harmonic mean of accuracy and recall is taken for the performance evaluation of systems. This paper explores twenty CWI systems and identifies that why sv000gg system outperformed with highest G-score as 0.773 and 0.774 for the two respective submissions.

Article History

Received: 27 May 2017
Accepted:23 June 2017

Keywords

CWI,
Lexical Simplification,
Textual Entailment,
Text Classification.

Introduction

The process of recognizing difficult words from text is the main task of Complex Word Identification (CWI). In this task, the difficult words are substituted with simpler words aim to enhancement of reader's understanding. CWI is useful in Lexical Simplification integrated with text simplification requires accurate identification of complex words from sentence. CWI can assist non native speakers by providing summarization of stories and generating simplified abstracts of essays. This can assist second language learners and dyslexic users through enhancement of understanding of complex web text. This way CWI is beneficial for naïve lerners too, making their lessons more readable by replacing difficult words with commonly used words.

SemEval-2016 Task-11 provided data set of 9200 sentences with word range of 20 to 40 words per sentence. This data set is generated from three sources as; CW corpus (Shardlow, 2013b), Lex M Turk Corpus (Horn. et. al. 2014) and Simple

CONTACT JASPREET SINGH PROFJASPREETBATTH@GMAIL.COM Research Scholar, Department of Computer Science, Guru Nanak Dev University Amritsar. Punjab, India

To link to this article: http://dx.doi.org/10.13005/ojcst/10.03.09

Wikipedia corpus by (Kauchak, 2013). The CW corpus contains 731 simple English sentences in which one complex word is substituted by Wikipedia editors. The second data set Lex M Turk is commonly used for CWI is composed of 500 sentences from Simple English and containing one complex word in each sentence. The third data set Simple Wikipedia composed of 1,67,689 sentences taken from simple English Wikipedia sources. The CWI systems are being leveraged on 8700 sentences from the third data set.

Consider a sentence as an example: "Our University follows the praxes of Guru Nanak Dev Ji". The automated system firstly collects complex words using complexity measures described in table-1 below, eg. praxes, then look for suitable matches that can be appropriately replaced with it without affecting the meaning of sentence. A thesaurus lookup produces the following replacements: practices, rehearse, exercise and drill. Here rehearse, exercise and drill need to be dropped because they are falling out of context. Finally the automated system would find appropriate substitute as 'practices' and substitute it with its complex variant, generating the simple sentence: "Our University follows the practices of Guru Nanak Dev Ji". The following Fig. 1 shows the lexical simplification of text through CWI.

Two types of techniques are used for CWI: Threshold based CWI and Classification based CWI. Threshold based CWI techniques are compared by (Shardlow, 2013) []. Corpus of complex words is collected from Wikipedia in which pairs of sentences (XwY, X Y) are extracted based on edit history and different annotations of word 'w' as complex. One of the pioneers in CWI is (Carroll et. al, 1998) [] considered word frequency as a parameter for CWI. Those words whose frequency is lying under some threshold value are considered as complex or uncommon words. Classification based techniques uses some machine learning algorithm eg. SVM to train a classifier which uses word features to decide the complexity of word. Word features that resemble the trained word's features are taken to be complex words. Following are the threshold based scales for measuring text complexity:

Related Work

Automated text simplification started in 1996 by (Chandersehar & Srinivas)[1], they have performed superficial analysis of text to identify noun and verb phrases from complex sentences. In[2], Siddharthan, 2006 concluded that lexical simplification is a subtask of text simplification in which complex phrases are substituted with wide eyed variants to enhance the readability of text. SemEval-2012 featured with text simplification tasks based on three

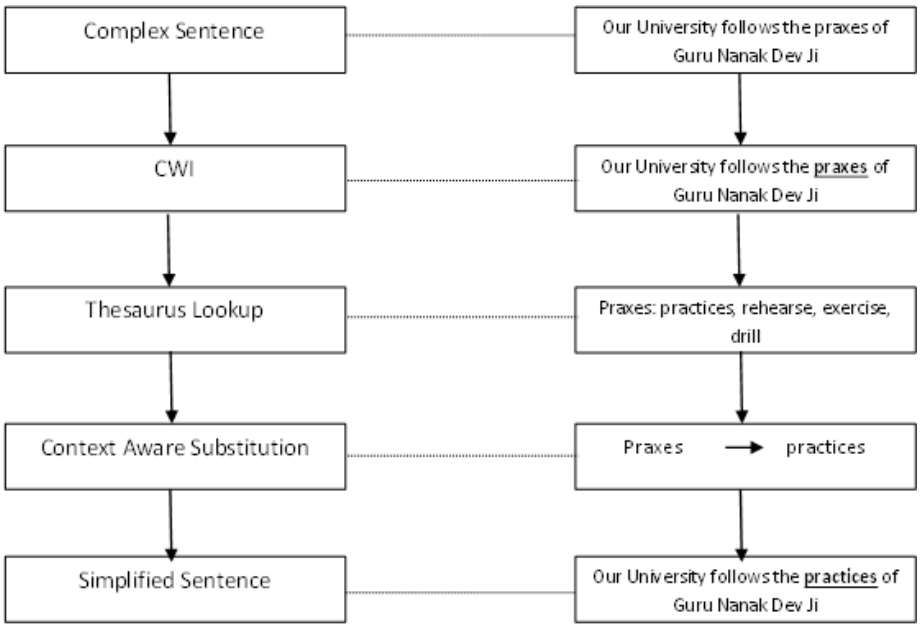

Fig. 1: Lexical Simplification process using CWI

aspects as complexity analysis, search for substitute words and ranking of substitute variants.

In 2010, Lucia Specia et. al.[3] described English text simplification using context aware lexical simplification approaches. They have outperformed with the best G-score among nine participating teams in SemEval-2012. In 2012, De Belder et. al.[4] proposed a method in which combination of two sources as lexicon and language model is introduced. Out of given text, the idea is to generate two lists; one list contains synonyms from lexical databases while second list holds alternative words generated through latent word models. A probabilistic model then estimates the probability of substitute for original complex word. In 2013, A. Di. Marco & R. Navigli[5] proposed a graph based Word Sense Induction (WSI) model for clustering and diversifying results of web search text. They have automated the task of WSI by evaluating semantic similarities from raw text and discovering words senses from them. In SemEval-2013, D. Jurgen & I.P. Klapaftis[6] measured the performance of Word Sense Disambiguation (WSD) systems by discovering a software bug in them. The bug was wrongly labelling word senses and resulting in a wrong interpretation of words. In SemEval-2014, M. Marelli et. al.[7] presented a model for finding semantic relatedness and textual entailment of English sentences. They have decomposed the dataset into two halves for training and testing the classifiers. Pair of sentences are taken in lexical entailment process and degree of relatedness is measured in terms of Pearson's correlation and Spearman's correlation coefficients. In the same competition 2014, S. Oepen et. al.[8] defined a Semantic Dependency Parsing (SDP) system for extracting internal structure of sentences by collecting predicate-argument pairs for context words. In SemEval-2015, E. Agirre et. al.[9] submitted systems for Semantic Textual Similarity (STS) to identify the degree of relatedness between two text snippets. In the same series of tasks in 2015, A. Moro & R. Navigli[10] presented a system for WSD and entity linking in multilingual texts. The idea of taking CWI in competitive series was conceived during SemEval-2016 Task11, H.P. Gustavo & Lucia Special[11] found CWI systems capable of identifying complex words from text and assisting lexical simplification of text. In SemEval-2016, Sanjay S.P et. al.[18] attained handsome accuracy in CWI using SVM linear classifier.

Cwi Systems At Semeval-2016 Task 11

SemEval is the semantic evaluation series of computational linguistics held in San Diego, California every year. SemEval-2016 has provided a platform for different linguistic tasks to the Linguistics-Computing professionals. One of the tasks was CWI in which total 42 teams across the world participated and 20 teams have submitted their systems for CWI[12].

Table 1: Threshold based scales for measuring complexity of text.

Scale	Proposed by	Mathematical Model	Description of variables	Range Indicator
Flesch Reading Ease	Flesch (1949)	S =206.835 (1.015*ASL)- (84.6*ASW)	S is Flesch's scale ASL is Average Sentence Length ASW is Average number of Syllables per word	(0-30) Complex (30-70) Medium (70-100) Easy
FOG Index	Gunning (1952)	GL= 0.4*(ASL + #of Hard Words)	GL is grade level	(0-6) Easy (6-12) Complex
Flesch-Kincaid	Kincaid *et. al.,* -1986	Text_Level (0.39*ASL)+ (11.8*ASW)- 15.59	Text Level Measures the complexity in text	(6-10) Easy (>10) Complex

Following is the summary of systems with their accuracy of findings. The evaluation is carried out in terms of G-score metric which is the Harmonic Mean of accuracy and recall.

Best Performed System Sv000gg

The reason behind the highest G-score of sv000gg system is the ensemble of machine learning classifiers. The system does hard voting of complex word labels predicted through different classifiers. The soft voting in which system classifies the estimates of complexity through maximum argument of traditional hard voting. This makes the system confident to classify the complex words in the given context from the text snippet.

The other reason for the performance of the system is features used in classification process. This system has used total 69 features which are grouped into four categories as Binary, Lexical, Collocational and Nominal features covering the wide range of features than other systems in the competition. In this system the training of classifiers is done through 21 voters which are grouped into three categories as Lexicon based, Threshold based and Machine learning. The third group of machine learning contains the ensemble of seven Machine learning algorithms due to which system got strengthened with more preciseness in classification. Further the results are undergone through five fold cross validation over a joint dataset.

Challenges In Text Simplification

One of the biggest challenges in Natural Language Processing is ambiguity problem. Since last decade many researchers have tried to reduce the ambiguity through word sense disambiguation techniques. CWI also comprises many challenges including ambiguity. Accuracy of appropriate substitutes depends on dataset used for classification based CWI. Results of text simplification may not be promising if a classifier is trained using immature dataset. In SemEval-2016, the submissions of twenty teams worked in the same direction to enhance the lexical simplification process of web text. They have faced so many challenges while developing solutions for CWI. Following are the challenges in CWI for lexical simplification of text:

1. To Accurately identify Complex Words
2. An appropriate thesaurus lookup (Avoid jargons and dyslexic phrases)
3. The Context aware substitution (Substitutes of complex variants should preserve the semantic structure of sentence)
4. To measure the degree of ambiguity of complex words (Word sense disambiguation)
5. To make non-native English user understand the complex challenging words in complex sentences.
6. To identify simplification needs of individuals by comparing complexity of words with overall users of English on the web of same category.
7. To build a new corpus to be used in Lexical Simplification and other tasks related to semantic evaluations.
8. To measure the applicability of different datasets used in formation of CWI systems.
9. To enhance the performance matrices of CWI systems for English text.
10. To investigate various word parameters used in Lexical Simplification process.

Conclusion

Automated CWI systems are quite useful in assisting aphasic users, non-native English users, second language learners and students as per their simplification needs. The investigation of twenty CWI systems is carried out on the basis of G-score measure of their performances. The SVG000GG system seems to be on the highest position in terms on the G-score. The main reason for the higher performance is due to coverage of wide range of Machine Learning classifiers along with more number of word features considered than other systems. The accuracy of the classification is also validated through five fold cross validation on the joint dataset. This system has brought competition among other systems to incorporate ensemble of classifiers into CWI process. The future aspect of CWI systems will leverage Deep Learning Techniques and Convolution Neural Networks for better performance.

References

1. Chandrasekar, R., Doran, C., & Srinivas, B. . Motivations and methods for text simplification. Proceedings of the 16th conference on Computational linguistics -. doi:10.3115/993268.993361, (1996).
2. Siddharthan, A. Syntactic Simplification and Text Cohesion. Research on Language and Computation, vol.**4**(1), pp.77-109. doi:10.1007/s11168-006-9011-1 (2006).
3. Specia, L. Translating from Complex to Simplified Sentences. *Lecture Notes in Computer Science*, doi:10.1007/978-3-642-12320-7_5, pp.30-39. (2010).
4. De Belder, J., & Moens, M. . A Dataset for the Evaluation of Lexical Simplification. Computational Linguistics and Intelligent Text Processing,doi:10.1007/978-3-642-28601-8_36, pp.426-437 (2012)
5. Di Marco, A., & Navigli, R. Clustering and Diversifying Web Search Results with Graph-Based Word Sense Induction. Computational Linguistics, doi:10.1162/coli_a_00148, vol.**39**(3), pp.709-754. (2013).
6. Klapaftis, I. P., & Manandhar, S. Evaluating Word Sense Induction and Disambiguation Methods. *Language Resources and Evaluation*, doi:10.1007/s10579-012-9205-0 vol.**47**(3), pp.579-605. (2013).
7. Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. SemEval -2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). doi:10.3115/v1/s14-2001 (2014).
8. Oepen, S., Kuhlmann, M., Miyao, Y., Zeman, D., Flickinger, D., Hajic, J., Zhang, Y. SemEval 2014 Task 8: Broad-Coverage Semantic Dependency Parsing. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). doi:10.3115/v1/s14-2008 (2014).
9. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., *Wiebe, J.* SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). doi:10.3115/v1/s14-2010 (2014).
10. Moro, A., & Navigli, R. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). doi:10.18653/v1/s15-2049 (2015).
11. Paetzold, G., & Specia, L. (2015). LEXenstein: A Framework for Lexical Simplification. Proceedings of ACL-IJCNLP 2015 System Demonstrations. doi:10.3115/v1/p15-4015
12. Paetzold, G., & Specia, L. SemEval 2016 Task 11: Complex Word Identification. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1085, (2016).
13. Davoodi, E., & Kosseim,L. CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic Features for Complex Word Identification. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1151 (2016).
14. Konkol, M. (2016). UWB at SemEval-2016 Task 11: Exploring Features for Complex Word Identification. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval- 2016). doi:10.18653/v1/s16-1162
15. Kuru, O. (2016). AI-KU at SemEval-2016 Task 11: Word Embeddings and Substring Features for Complex Word Identification. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1163
16. Martínez Martínez, J. M., & Tan, L. USAAR at SemEval-2016 Task 11: Complex Word Identification with Sense Entropy and Sentence Perplexity. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1147, (2016).
17. Paetzold, G., & Specia, L. SV000gg at SemEval-2016 Task 11: Heavy Gauge

Complex Word Identification with System Voting. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval -2016). doi:10.18653/v1/s16-1149 (2016).

18. Sp, S., Kumar,A., & K P, S. AmritaCEN at SemEval-2016 Task 11: Complex Word Identification using Word Embedding. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1159 (2016).
19. Wróbel, K. PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval- 2016). doi:10.18653/v1/s16-1146, (2016).
20. Choubey, P., & Pateria, S. Garuda and Bhasha at SemEval-2016 Task 11: Complex Word Identification Using Aggregated Learning Models. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1156, (2016).
21. Malmasi, S., Dras, M., & Zampieri, M. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). doi:10.18653/v1/s16-1154, (2016).