New Data is Indeed Helping Lexical Simplification

by

Ashish Plakurthi, Radhika Mamidi

in

18th International Conference on Computational Linguistics and Intelligent Text Processing (*CICLing-2017*)

Report No: IIIT/TR/2017/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
April 2017

New Data is Indeed Helping Lexical Simplification

Ashish Palakurthi and **Radhika Mamidi**
Language Technologies Research Center
KCIS, IIIT-Hyderabad, India

ashish.palakurthi@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract. We propose the use of the Newsela corpus for Complex Word Identification, a sub-problem of Lexical Simplification and conduct an empirical evaluation by comparing it with benchmark corpora previously employed for this task. Our experiments suggest that the proposed corpus is effective for Complex Word Identification, thus helping Lexical Simplification.

1 Introduction

Lexical Simplification (LS) is a procedure focused to enhance the readability of the given text by transforming complex text into simple text. LS [1], [2],[3] [4] is the method of substituting a word in a given context with its best (easiest to understand) substitute that can improve the readability of the text. It is critical to ensure that the meaning of the text is preserved while choosing the new substitute and replacing the target word. LS [5] provides a gamut of applications beneficial to audiences like people with aphasia, children and most importantly non-native speakers.

LS in general can be divided into four sub-tasks:

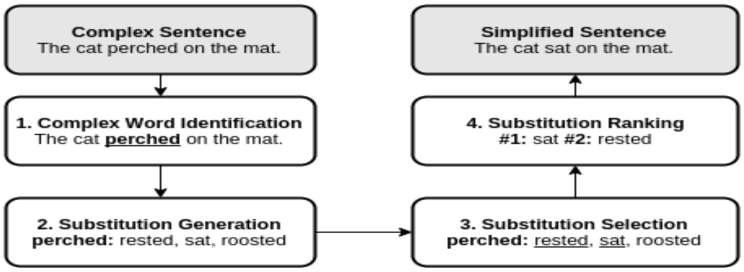
Complex Sentence
The cat perched on the mat.
1. Complex Word Identification
The cat perched on the mat.
2. Substitution Generation
perched: rested, sat, roosted
3. Substitution Selection
perched: rested, sat, roosted
4. Substitution Ranking
#1: sat #2: rested
Simplified Sentence
The cat sat on the mat.

Fig. 1. Lexical Simplification Pipeline (Image reproduced from [6])

- *Complex Word Identification (CWI):* This is the first step in the pipeline of Lexical Simplification. It is the process of identifying complex[1] words in a sentence which are to be simplified.
- *Substitute Generation:* Substitute generation is the process of producing the variants of the words identified in the previous step.
- *Substitute Selection:* Selecting the right variants which fit the context of the sentence is the main aim of this sub-task, as it is imperative to preserve the meaning of the sentence.
- *Substitute Ranking:* This sub-task ranks the words identified in the previous step, picks the simplest word and then replaces the complex word identified in the first step with this simplest word.

As seen in Fig. 1, the overall performance of a Lexical Simplification system is crucially dependent upon CWI as it is the first step in the flow.

Recent research [2] has shown that easily readable corpora can be helpful in building better LS systems. In this paper, we explore the usefulness of the Newsela corpus [7] for step 1 (Figure 1), CWI for non-native speakers. In particular, we analyze its empirical performance and compare the Newsela corpus with other benchmark corpora commonly used for LS and previously employed for CWI.

Our motivations behind exploring the Newsela corpus for CWI are two-fold:

- One, Newsela particularly aims to ease text readability for children studying at different grade levels by providing easy-to-understand articles, written by professionals. This is compatible with the goal of LS [2], where the authors show the importance of tackling LS with respect to target audience like non-native speakers. The authors also show that the vocabulary of non-native speakers correlates well with that of children .
- Two, recent work [7] confirms that the Newsela's quality of simplification is far better than Simple Wikipedia which is regularly used for LS.

Our experiments suggest that in comparison to commonly used corpora like Simple Wikipedia and SubIMDB, the Newsela corpus is more beneficial for CWI and can thus aid in improving LS for non-native English speakers.

2 Related Work

CWI has emerged as a new challenge in the field of Natural Language Processing, where only a few approaches have been attempted in addressing this task. A recent work gauges simplicity score [8] of a word by integrating length of a word along with its frequency in a corpus. A threshold value T is chosen, where a word is simplified only if the word's frequency is lower than the fixed threshold T. Matthew Shardlow [9] explored a similar approach to differentiate between simple and complex words, where experiments were carried with each threshold value on a particular corpus. In addition, there have been various machine

[1] complex: difficult to understand.

learning based methods [10] attempted to distinguish simple words from complex words.

Substitution generation, the task of finding a set of candidate substitutes for the target complex word, is a task for which linguistic databases were queried [11], [12]. A few works focused on paraphrase-based methods [13], [14] and other works [15], [4] depended on parallel sentences from Wikipedia and Simple Wikipedia to generate candidate variants.

Substitution Selection, where the goal is to find the word that best replaces the complex word and also make sure that it fits well in the context of the target word to be replaced, has been relatively explored to a greater extent. Word Sense Disambiguation methods [16], [17] have been investigated, where substitutions which did not share a similar sense with the target words were discarded. Apart from this, to differentiate ambiguous words, POS tags of words were also utilized [18]. More interestingly, semantic similarity methods [19] measured word similarity using cosine distance between context vectors of the words to find best substitutions for a target word.

The Substitution Ranking task, involving to choose the simplest candidate from a set of candidates generated in the previous step (Substitution selection), has also been studied through a variety of approaches. Frequency of a word [1] in a corpus was found to be helpful in ranking the substitutes. Ranking candidates based on length and frequency was also found to be efficient [3] for this sub-task.

The remainder of this paper is structured as follows. In Section 3, we describe the data used for our experiments. We also discuss interesting insights and show results in the same section. We conclude in Section 4.

3 Experiments and Discussions

3.1 Data

We mainly use two corpora for our experiments:

- **Newsela corpus:** We use the Newsela corpus to investigate its usefulness for CWI task. The Newsela corpus consists of 4 simplified versions (Simp-1,2,3,4) of News articles, where Simp-1 is the least simplified version and Simp-4 is the most simplified version [7]. Each version comprises of various articles, where the articles are easy or difficult to comprehend based on the simplicity version. Table 1 shows the corpus statistics of each version.

- **SemEval 2016 CWI corpus:** For evaluating our experiments, we use the SemEval-2016 Complex Word Identification shared task [10] corpus, comprising of 2237 training data instances and 88,221 test data instances where each instance is of the form :
  $<sentence><word><position><complexity>$
  where *sentence* is the given sentence, *word* is the target word in the sentence whose $<complexity>$can be 1 if the target word is complex (difficult) and 0 if the target word is simple. The *position* field is the position of the target

Version	No. of articles	No. of Tokens
Simp-1	1910	1896826
Simp-2	1910	1746938
Simp-3	1910	1474253
Simp-4	1882	1142330

Table 1. Newsela corpus statistics

word in the given sentence. Table 2 shows statistics of the SemEval-2106 CWI corpus.

	Instances	Simple	Complex
Train	2237	1531	706
Test	88221	84090	4131

Table 2. SemEval corpus statistics

One can clearly notice that only 4% of the test set contains the complex words, which made the CWI Shared Task [10] very challenging [10]. According to the CWI shared task organizers[10], the data was collected and annotated by 400 non-native speakers, wherein each participant marked the word as simple if they could understand the word, else they marked it as complex. After everyone finished their annotation, for every instance in the experiment, the final tag assigned to a word was complex even if one annotator marked had the word as complex.

Examples taken from the training corpus of CWI shared task [10]:

- In *A frenulum is a small fold of tissue that secures or restricts the motion of a mobile organ in the body,* the target word *frenulum* was deemed as complex (by a non-native speaker).
- In *It resembles five deep spoons with the handles linked , or , alternately , the hammocks resemble five fig halves .*, the target word *hammocks* was deemed as complex (by a non-native speaker).

Apart from the above, other state-of-the-art corpora used for comparisons in our experiments include the SubIMDB [6], which is a corpus containing SubIMDB movie subtitles. Ogden's dictionary[20] which is a list of basic and commonly used English words. SUBTLEX [21] is another corpus containing subtitles of films and television series. Other corpora used are the Wikipedia and Simple Wikipedia [22].

3.2 Evaluation

G-score, a harmonic mean of accuracy and recall, the official evaluation metric of the Shared Task, is mainly used to evaluate experimental results. G-score helps in rewarding systems that are good in predicting the complexity of not only complex words (Precision), but rather all words (Accuracy).

$$G - score = \frac{2 * A * R}{A + R}$$

3.3 Experiments

To demonstrate the usefulness of a corpus, we conduct different experiments. Our experiments incline towards methods that best reveal the usefulness of a corpus for CWI. Word tokenization for experiments is done using tokenization script previously [7] used for Newsela corpus. The experiments are:

1. **Lexicon Based** (E_1) [10],[23]: Target word is considered as simple if it is present in a vocabulary list, else it is complex. All words in the Newsela corpus are taken as a vocabulary list. Similary, we used different vocabulary lists for other corpora.
2. **Threshold Based** (E_2) [10], [23]: We compute the unigram probability (over a corpus) of each target word in the training data, find the minimum and maximum values, divide the interval between minimum and maximum into 10,000 equal parts, perform a search over all the 10,000 values and find the threshold value T that gives the highest G-score on training data. All target words in the test data with unigram probabilities greater than T are then labeled as simple, while the remaining words are labeled as complex.
3. **Simplicity** (E_3): To verify the simplicity of different versions of the Newsela corpus, we estimate how frequently simple words occur with respect to complex words in a corpus. These simple and complex words are taken from CWI Shared Task test data. We compute a simplicity score, *S(C)* for each corpus *C*. *S(C)* is defined as:

   $$S(C) = \frac{\sum_{i=1}^{n_s} f(w_{s_i})}{\sum_{i=1}^{n_c} f(w_{c_i})}$$

   where n_s, n_c are the total number of simple and complex words respectively, w_{s_i} and w_{c_i} represent i^{th} simple and i^{th} complex word respectively and *f(w)* represents the frequency of a word w in a corpus C. A higher simplicity score indicates that the corpus is easier to comprehend than a corpus with lower simplicity score.

3.4 Results

Table 3 shows results obtained in experiment E_1 and Table 4 shows results obtained in experiment E_2 for CWI. In both the tables, A is the accuracy, P is the precision, R is the recall, F is the F-score and G is the G-score.

In experiment E_1, all Simp-* corpora significantly beat other benchmark corpora, when evaluated over both G-score and F-score. Simp-4 achieves a G-score of 0.713 in and ranks 11^{th}(out of 42) if compared with other systems in the Shared Task [10]. Without any optimization on F-score, the Simp-2 corpus achieves an F-score of 0.343 (2nd best if compared to other systems in the Shared Task), while the best performing system, optimized on F-score, achieved an F-score of 0.353. This is remarkable for a simple lexicon based system.

System	A	P	R	F	G
SubIMDB	0.913	0.217	0.332	0.262	0.487
Ogden's	0.248	0.056	0.947	0.105	0.393
Wikipedia	0.047	0.047	1.000	0.089	0.090
Simple Wikipedia	0.953	0.241	0.002	0.003	0.003
Simp-1	0.926	0.291	0.402	0.338	0.561
Simp-2	0.917	0.273	0.462	**0.343**	0.614
Simp-3	0.894	0.233	0.545	0.326	0.677
Simp-4	0.864	0.195	0.607	0.295	**0.713**

Table 3. Lexicon Based Results

In experiment E_2, the Simp-1 corpus achieves the highest G-score and F-score (Table 4) in comparison to other benchmark corpora. However, the reason for Simp-1 performing better than the more simplified versions like Simp-4 could possibly be due to the larger size [1] of Simp-1 corpus in comparison to Simp-4. This is observed even in the case of Wikipedia and Simple Wikipedia with Wikipedia having higher number of tokens.

Table 5 shows results obtained in experiment E_3, where S denotes the Simplicity scores obtained for each corpus. As can be seen from Table 5, Simp-4 achieved the highest Simplicity score and Simp-1 obtained the lowest score among all the four Newsela versions.

3.5 Discussions

For CWI, it is the complexity (complex or simple) of words in the corpus that plays a major role in achieving better results as seen in Experiments E_1 and E_3. This property is observed within the Newsela corpora of different simplicity levels, and also observed in case of Wikipedia and Simple Wikipedia.

System	A	P	R	F	G
SubIMDB	0.445	0.072	0.912	0.133	0.598
SUBTLEX	0.492	0.077	0.896	0.142	0.636
Wikipedia	0.536	0.084	0.901	0.154	0.672
Simple Wikipedia	0.513	0.081	0.902	0.148	0.654
Simp-1	0.579	0.090	0.884	**0.164**	**0.699**
Simp-2	0.546	0.085	0.895	0.156	0.678
Simp-3	0.524	0.082	0.901	0.150	0.663
Simp-4	0.506	0.079	0.904	0.146	0.649

Table 4. Threshold Based Results

System	S
Simp-1	137.69
Simp-2	139.26
Simp-3	141.28
Simp-4	143.13

Table 5. Simplicity Scores

Apart from the promising scores seen in the previous section, few interesting findings regarding the Newsela corpus are:

- While the Newsela corpus is predominantly targeted for children, the evaluation data developed by the SemEval-2016 CWI Shared Task organizers reflects only non-native speakers above the age of 18 years [10]. The results (Table 3, Table 4) obtained using the Newsela corpus to an extent reveal that vocabulary difficulties faced by non-native speakers are reasonably similar to that of children.
- The SubIMDB corpus, a huge subtitles-compilation of movies targeting children, does not handle vocabulary difficulties of non-natives as effectively (Table 3, Table 4) as the Newsela corpus (which also primarily targets children) does.
- The Simplicity scores obtained in experiment E_3 correlate with the simplicity versions [7] of the Newsela corpus. S(Simp-1): 137.69, S(Simp-2): 139.26, S(Simp-3): 141.28 and S(Simp-4): 143.13 as shown in Table 5. A monotonic increase in simplicity scores is observed. Clearly, these scores indicate that simple words for non-natives occur more frequently with respect to complex words in Simp-4 than in Simp-1.
- The G-scores (Table 3) in E_1 monotonically increase from Simp-1 to Simp-4 and correlate with the simplicity versions of the Newsela corpus. In addition, the G-score of Simple Wikipedia in E_1 is lower than all Newsela versions.

This is compatible with previous quantitative findings [7], where the authors claim that the words in the Newsela corpus are shorter on an average, thus simpler than the words in Simple Wikipedia.

The potentiality to consistently differentiate vocabulary across multiple levels is a striking property that distinguishes the Newsela corpus from benchmark corpora like Simple Wikipedia, SUBTLEX, SubIMDB etc. which can be more helpful for non-native English speakers.

In contrast to E_1, experiment E_2 shows a monotonic decrease of G-scores from Simp-4 to Simp-1. Interestingly, the SubIMDB corpus with a G-score higher than Wikipedia and Simple Wikipedia in E_1 ends up with the least G-score in E_2. This could possibly be due to the unconventional train-test split, given that the threshold T in E_2 is dependent on the training data. Adding to that, the class distribution in both training and test data is highly imbalanced. However, we look to closely investigate the reason behind this phenomenon.

4 Conclusion

Our experiments empirically reaffirm the claims of previous work [7], showing that the Newsela corpus can help LS better. More importantly, experimental evaluation was done using data [10] collected from non-native speakers. We hope that our findings encourage the research community towards using the Newsela corpus for LS. In future work, we plan to explore the utility of this corpus for steps 2,3 and 4 in the LS pipeline shown in Figure 1.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive insights. We are grateful to Luke Orland for providing us the Newsela data.

References

1. Specia, L., Jauhar, S.K., Mihalcea, R.: Semeval-2012 task 1: English lexical simplification. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics (2012) 347–355
2. Paetzold, G.H., Specia, L.: Unsupervised lexical simplification for non-native speakers. In: Thirtieth AAAI Conference on Artificial Intelligence. (2016)
3. Jauhar, S.K., Specia, L.: Uow-shef: Simplex–lexical simplicity ranking based on contextual and psycholinguistic features. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics (2012) 477–481

4. Horn, C., Manduca, C., Kauchak, D.: Learning a lexical simplifier using wikipedia. In: ACL (2). (2014) 458–463
5. Siddharthan, A.: A survey of research on text simplification. ITL-International Journal of Applied Linguistics **165** (2014) 259–298
6. Paetzold, G.H.: Reliable lexical simplification for non-native speakers. In: NAACL-HLT 2015 Student Research Workshop (SRW). (2015) 9
7. Xu, W., Callison-Burch, C., Napoles, C.: Problems in current text simplification research: New data can help. Transactions of the Association for Computational Linguistics **3** (2015) 283–297
8. HoracioSAGGION, S.L.B.: (Can spanish be simpler? lexsis: Lexical simplification for spanish)
9. Shardlow, M.: A comparison of techniques to automatically identify complex words. In: ACL (Student Research Workshop), Citeseer (2013) 103–109
10. Paetzold, G.H., Specia, L.: Semeval 2016 task 11: Complex word identification. Proceedings of SemEval (2016) 560–569
11. Sinha, R.: Unt-simprank: Systems for lexical simplification ranking. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics (2012) 493–496
12. Chen, H.B., Huang, H.H., Chen, H.H., Tan, C.T.: A simplification-translation-restoration framework for cross-domain smt applications. In: COLING. (2012) 545–560
13. Elhadad, N., Sutaria, K.: Mining a lexicon of technical terms and lay equivalents. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Association for Computational Linguistics (2007) 49–56
14. Kauchak, D., Barzilay, R.: Paraphrasing for automatic evaluation. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics (2006) 455–462
15. Feblowitz, D., Kauchak, D.: Sentence simplification as tree transduction. In: Proc. of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations, sn (2013) 1–10
16. Sedding, J., Kazakov, D.: Wordnet-based text document clustering. In: proceedings of the 3rd workshop on robust methods in analysis of natural language data, Association for Computational Linguistics (2004) 104–113
17. Nunes, B.P., Kawase, R., Siehndel, P., Casanova, M.A., Dietze, S.: As simple as it gets-a sentence simplifier for different learning levels and contexts. In: Advanced Learning Technologies (ICALT), 2013 IEEE 13th International Conference on, IEEE (2013) 128–132
18. Aluísio, S.M., Gasperin, C.: Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In: Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, Association for Computational Linguistics (2010) 46–53
19. Biran, O., Brody, S., Elhadad, N.: Putting it simply: a context-aware approach to lexical simplification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, Association for Computational Linguistics (2011) 496–501

20. Ogden, C.K.: Basic English: international second language. Harcourt, Brace & World (1968)
21. Brysbaert, M., New, B.: 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** (2009) 977–990
22. Kauchak, D.: Improving text simplification language modeling using unsimplified text data. In: ACL (1). (2013) 1537–1546
23. Paetzold, G.H., Specia, L.: Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. Proceedings of SemEval (2016) 969–974