CEFR-based Lexical Simplification Dataset

Satoru Uchida[†], Shohei Takada[‡], Yuki Arase[‡]
[†] Faculty of Languages and Cultures, Kyushu University
[‡] Graduate School of Information Science and Technology, Osaka University
[†] uchida@flc.kyushu-u.ac.jp, [‡] {takada.syouhei, arase}@ist.osaka-u.ac.jp

Abstract

This study creates a language dataset for lexical simplification based on Common European Framework of References for Languages (CEFR) levels (CEFR-LS). Lexical simplification has continued to be one of the important tasks for language learning and education. There are several language resources for lexical simplification that are available for generating rules and creating simplifiers using machine learning. However, these resources are not tailored to language education with word levels and lists of candidates tending to be subjective. Different from these, the present study constructs a CEFR-LS whose target and candidate words are assigned CEFR levels using CEFR-J wordlists and English Vocabulary Profile, and candidates are selected using an online thesaurus. Since CEFR is widely used around the world, using CEFR levels makes it possible to apply a simplification method based on our dataset to language education directly. CEFR-LS currently includes 406 targets and 4912 candidates. To evaluate the validity of CEFR-LS for machine learning, two basic models are employed for selecting candidates and the results are presented as a reference for future users of the dataset.

Keywords: Lexical simplification, ESL learners, paraphrasing, CEFR

1. Introduction

There is no doubt that vocabulary is the key to successful international communication. Laufer (1989) points out that learners of a foreign language need to know 95% of the words in the input text to be able to successfully understand the text message, which implies the importance of vocabulary in language learning. However, authentic English passages may contain difficult words that may hinder the readers' or listeners' comprehension.

One solution is to simplify words in the input text using statistical and computer-based methods (Horn et al., 2014; Biran et al., 2011; Glavas and Stajner, 2015). This task is called *lexical simplification* whose aim is to replace a difficult word (referred to as the *target* word in this paper) with a simpler word selected from candidates (referred to as *candidate* words in this paper), which has been featured at SemEval2012[1]. Although previous studies have shown that data-driven approaches to find a simpler word are useful to some extent (Horn et al., 2014; Glavas and Stajner, 2015), they do not place the focus on educational aspects and hence tend to discuss only technical issues. Also, the criteria for the word difficulty tend to be vague and subjective, which would imply that previous research methods may not be useful for suggesting simplifications for different proficiency levels of learners.

The present study attempts to construct a dataset that uses Common European Framework of Reference for Languages (CEFR) as a criterion for word levels using introductory parts of university textbooks whose contents are academically reliable, educationally important, and covering a variety of topics. In this sense, our dataset is education-oriented based on a solid and widely-used framework for language education. CEFR levels are assigned to both the target and candidate words for lexical simplification. This allows us to adjust the text level flexibly according to learners' proficiency, which is especially meaningful for educational purposes. Also, two basic models for lexical simplification will be applied to our dataset in order to show the performances of baseline methods.

Food is procured with its suckers and then crushed using its tough "beak" of chitin.

Target: procured
Candidates (frequency): obtained (17), gathered (9), gotten (8), grabbed (4), ...

Table 1: Horn et al. (2014) example 1

2. Related Work

Coster and Kauchak (2011b) created and published a corpus for *text* simplification pairing sentences from Wikipedia and Simple Wikipedia (hereafter referred to as the Wikipedia corpus). This dataset has 137K pairs of simplified and unsimplified sentences that can be used for creating simplification rules. Coster and Kauchak (2011a) was one such attempt, and they examined a variety of paraphrasing rules including lexical changes, reordering, insertions and deletions using the Wikipedia corpus. Another attempt was Horn et al. (2014), who evaluated paraphrases using language models based on several language resources.

For evaluation tasks of lexical simplification, Horn et al. (2014)[2] published a dataset consisting of 500 sentences from the Wikipedia corpus with the results of annotations by Amazon Mechanical Turk[3]. In this dataset, they asked 50 turkers for each sentence to replace a target with a simpler one, and ranked the candidate according to the frequency (see Table 1) . The results seem to be natural and intuitive, but the levels of the candidates are not necessarily easy for learners because of the lack of educational criteria.

[1] https://www.cs.york.ac.uk/semeval-2012/task1.html

[2] http://www.cs.pomona.edu/~dkauchak/simplification/

[3] https://www.mturk.com/mturk/welcome

The pulses were either short or long, representing the dots and dashes of Morse code.

Target: pulse
Candidates (ranking): sound (1st), beat (2nd), beep (3rd), pulse (4th), emanation (5th)

Table 2: SemEval 2012 dataset

But critics note that Francis Galton did not advocate coercion when he defined the principles of eugenics.

Target: advocate
Candidates: like, cause, apply, speak, allow, want, ...

Table 3: Horn et al. (2014) example 2

Another dataset to be mentioned here is task 1 of SemEval 2012 which asked researchers to rank the candidates in order of simplicity (see Table 2). The sentences were taken from task 10 of SemEval 2007[4], where candidates were given by five English native speakers and ranked by non-native speakers of English. Although this may possibly reflect learners' intuition about the difficulty of words, it is hard to say that the judgments are consistent because of the differences in their mastery of English.
It should be also pointed out that there are some words in the list of candidates that require outside knowledge for replacement. Table 3 from the dataset of Horn et al. (2014) serves as an example. If we take the verb "advocate" out of the context, it is difficult to list "like", "cause", and "apply" for substitution. In this sense, the candidate lists are not consistent, which may make the task using the dataset beyond the scope of lexical simplification.

3. CEFR-LS

To overcome the disadvantages of the existing datasets, this study created a CEFR-based language resource ver. 1.0 (hereafter referred to as CEFR-LS). The input sentences were taken from introductory chapters of university textbook available at OpenStax website[5] by Rice University[6]. One of the reasons to use university textbooks is that they are deemed to be excellent in quality and their contents are reliable and well-organized. Also, the OpenStax website provides introductory books on a number of academic fields, and hence the dataset can be extended to cover various topics. Finally, introductory textbooks are literally a gateway to academic fields, and thus it is meaningful to provide an assistance in lexical aspects of these textbooks.

3.1. Procedure

Some introductory textbooks on the OpenStax website were randomly selected (the dataset includes those of economics, psychology, sociology and so on) and the introduction parts were used for setting up the dataset.

Economics is the study of how humans make decisions in the face of scarcity.

Target: scarcity (C2)
Candidates: *dearth* (NA), lack (A2), *paucity* (NA), shortage (B1), ...

Table 4: CEFR-LS example (words in italics are excluded)

Using CEFR-J wordlist (Tono, 2016) and English Vocabulary Profile (Cambridge University Press, 2015), a wordlist for CEFR levels was created. This list also contains part-of-speech information, which are useful to determine the level of a word in different part-of-speech (e.g. a word "address" is A1 when it is used as a noun, while B1 as a verb).
CEFR levels consist of A1 (elementary), A2, B1, B2, C1, and C2 (advanced). In our dataset, words that are ranked as B2, C1, and C2 are selected as target to be simplified.
For selecting candidates, an online thesaurus[7] was employed and words equal to or higher than B2 level were excluded. Table 4 shows an example in CEFR-LS. Synonyms for "scarcity" (C2) are "dearth" (NA), "lack" (A2), "paucity" (NA), "shortage" (B1) and so on, but only "lack" (A2) and "shortage" (B1) are listed as candidate in our dataset (NA represents that a candidate is not included in the CEFR wordlist).
Then, the paraphrasability was checked by a native speaker of English (male, late 30s), who has taught in Japan for more than 10 years. The dataset is still under development, but currently it includes 406 targets and 4912 candidates.

3.2. Annotation Standard

To make the judgments objective, we set up a guideline for annotation, which can also be helpful for the future updates of our dataset accompanied with additional judgments by several annotators. The procedure of checking paraphrasability consists of three stages: grammatical reformation stage, definition stage, and context stage.
Grammatical reformation is defined as addition or removal of words to maintain the coherence of the sentence. Candidates that fail at this stage will be unable to complete the sentence without need for the sentence to be reformed, regardless of their semantic proximity to the target word. Note that the morphology of the target (third person singular, past tense etc.) is considered to be applied to the candidate automatically in the process of paraphrasing. The following case serves as an example:

(1) Chemical engineering, materials science, and nanotechnology combine chemical principles and empirical findings to produce useful substances, ranging from gasoline to fabrics to electronic.
Target: range (B2)
Candidate: cover (A2)

In this sentence, the verb "range" is the target to be

[4] `http://nlp.cs.swarthmore.edu/semeval/tasks/task10/summary.shtml`

[5] `http://cnx.org/`

[6] Those books were created under Creative Commons 4.0 License (`https://creativecommons.org/`) and can be reused and distributed for research purposes.

[7] `http://www.thesaurus.com/`

paraphrased. The candidate "cover" does not fit due to the need to remove the preposition "from" to make the sentence grammatically correct.

At the definition stage, the semantic aspects of both the target and candidate words are taken into consideration. Candidates will be judged as semantically proximate if they paraphrase a portion of the semantics of the target and they provide the reader with a word they are more likely to be familiar with (an example will be shown later).

If a candidate passes these stages, the word is examined in the context stage, where each candidate is checked if it successfully conveys the nuance of the target word in the specific context and does not affect the meaning of a sentence. Here the register of a word is not considered except in extreme cases – simplification often results in candidates being simpler and therefore more informal words than the target words.

(2) This civic engagement ensures that representative democracy will continue to flourish and that people will continue to influence government.
Target: influence (B2)
Candidates: determine (B1), impress (A2), change (A1)

In (2), where the target is the verb "influence" (B2), "determine" (B1) passes the grammatical reformation stage but fails at the definition stage because it does not cover the semantics of "influence" in the meaning of affecting the way someone thinks or behaves, though it might do so in another context where determination causes a change in someone's way of thinking. On the other hand, the word "impress" (A2) passes this stage because it can be used to mean to affect the way someone believes. However, it does not pass the context stage because "impressing government" is not synonymous to "influencing government" in this context. Finally, the candidate "change" (A1) passes all the stages because it is grammatically appropriate, semantically proximate and contextually synonymous to the target word.

3.3. Characteristics of CEFR-LS

One of the characteristics of our dataset is that it is based on CEFR. In the field of language education, CEFR has become one of the most widely-used criteria for evaluating language ability. This means that it is possible to share the standard of word levels among many language learners and teachers. Furthermore, it enables us to connect lexical aspects with other language aspects such as grammar[8] based on the framework of CEFR. Also, the word level can be adapted flexibly based on a learners' proficiency. If a researcher wants to establish a system that simplifies words into A levels, then he/she can omit B1 level candidates from the list.

Table 5 shows the distribution of CEFR levels for both target and candidate words in SemEval2012, Horn et al. (2014), and CFER-LS. It is clear that other two resources contain a number of A1, A2, and B1 level words as a target that might not be necessary to be replaced, and B2, C1, and C2 level words as candidate that might still be difficult for learners.

	SemEval2012		Horn2014		CEFR-LS	
	TAR	CAN	TAR	CAN	TAR	CAN
Total	1710	8596	500	5010	406	4912
A1	616	1528	16	1119	0	1127
A2	414	1429	73	923	0	1495
B1	335	1762	152	1077	0	2290
B2	224	1080	95	493	301	0
C1	20	156	14	65	35	0
C2	20	147	18	64	70	0
NA	81	2494	132	1269	0	0
TAR ($\geq$ B2)	15.4%		25.4%		100%	
CAN ($\leq$ B1)	54.9%		62.9%		100%	

Table 5: CEFR levels of target and candidate words, where "TAR" and "CAN" abbreviate "target" and "candidate," respectively.

	SemEval2012	Horn2014
Number of targets (Coverage in the thesaurus as a headword)	1710 (94.3%)	500 (94.8%)
Number of candidates (Coverage in the thesaurus)	8596 (40.5%)	5010 (36.5%)

Table 6: Coverage of candidates in the thesaurus

	CEFR-LS
Number of targets	406
Number of candidates	4912
Number of correct candidates	961
Average Number of candidates per target	12.1
Average Number of correct candidates per target	2.4
Number of sentences	271
Average Number of words per sentence	23.0

Table 7: Detailed statistics of CEFR-LS

Another feature is that candidate words in CEFR-LS are all taken from a thesaurus. This means that most simplifications in our dataset do not require the world knowledge. This makes the simplification task more feasible and consistent. Table 6 shows the percentages of candidate words in each study that appear in the thesaurus via target words as a headword. It is obvious that less than half of the words are not included in the dictionary suggesting that outside knowledge is required for replacement in many cases.

Finally, our dataset contains not only correct candidates but also incorrect candidates in the list (see Table 7 for details). This allows researchers to examine distinctive statistical scores between correct and incorrect ones.

[8] http://www.englishprofile.org/english-grammar-profile

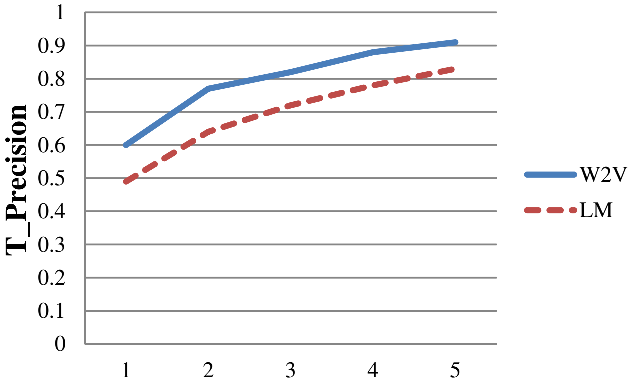

Figure 1: T_Precision curve for LM and W2V against top n

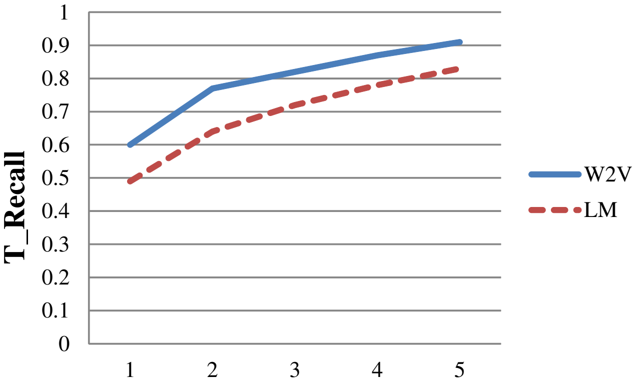

Figure 2: T_Recall curve for LM and W2V against top n

4. Performance of Baseline Methods

There have been several attempts in lexical simplification for selecting the correct candidate using statistical and machine learning methods such as Support Vector Machines (Horn et al., 2014), neural networks (Paetzold and Specia, 2017), and computation on word vectors (Glavas and Stajner, 2015). Also, we proposed (Takada et al., 2017) an approach using the collocation scores of target and candidate words. CEFR-LS is useful for evaluating these methodologies proposed in previous research.

As a preliminary attempt, this study uses two basic systems for selecting correct candidates to show the usefulness and validity of CEFR-LS[9]. One method is based on Horn et al. (2014) which uses a language model approach for candidate selection (hereafter LM). In this study, we constructed a language model using Google N-gram[10], and ranked candidates according to the language model probabilities of a sentence replaced a target with a candidate. The other method employed Word2Vec (Mikolov et al., 2013a; Mikolov et al., 2013b) and calculated the cosine similarity of candidates' vectors against targets' vectors (hereafter W2V), and candidates were arranged in order of similarity scores.

Precision and Recall were calculated with regard to the top n words in each model. Since finding one correct candidate for a target satisfies our purpose and the number of correct answers varies among targets, we adjusted these metrics to evaluate if at least a correct candidate is contained in the top n. We tentatively call them target-based precision and recall, represented as T_Precision and T_Recall in the paper, which are formally defined as follows.

$$\text{T\_Precision} = \frac{C^{\text{tar}}_{\text{correct}}}{C^{\text{tar}}_{\text{out}}},$$

$$\text{T\_Recall} = \frac{C^{\text{tar}}_{\text{correct}}}{C^{\text{tar}}},$$

where $C^{\text{tar}}_{\text{correct}}$ is the number of targets that are assigned at least a correct candidate, $C^{\text{tar}}_{\text{out}}$ is the number of targets that are assigned any candidates, C^{tar} is the number of all targets. Figure 1 shows the T_Precision curve and Figure 2 shows the T_Recall curve when varying the value of n. It turned out that LM and W2V have about 80% and 90% of T_Precision within the top five candidates respectively, which means a large portion of the targets that are assigned any candidates of CEFR-LS have at least one correct candidate in the top five words. However, it can be said that W2V outperforms LM in that it finds a correct answer more quickly (60% T_Precision for the top-ranked word).

On the other hand, LM has about 80% and W2V has about 90% T_Recall within the top five candidates. This result shows that 80% and 90% of the targets are assigned at least one correct candidate in the top five candidates, respectively.

There were 14 targets where both LM and W2V could not provide a correct paraphrase within the top 5 candidates. Among them are nine types of words including the following example:

(3) Division and specialization of labor only work when individuals can purchase what they do not produce in markets.
Target: division (B2)
Candidate: distribution (B1)

It is beyond the scope of the present paper to build a better algorithm for finding candidates, but we believe CEFR-LS would be a useful resource for this purpose.

5. Conclusion and Future Work

This study has shown that CEFR-LS is an education-oriented dataset for lexical simplification. It specifies CEFR levels for both target and candidate words, which makes the dataset consistent and especially relevant for language education. Also, candidates were selected using a thesaurus, making the simplification task feasible. The performance of two baseline methods on the CEFR-LS also provides a reference for future users of the dataset.

The current version of CEFR-LS is mainly intended as a pilot for constructing a language resource. Therefore, there remains room for future development. Clearly, one important task will be to increase the number of target words. It is planned to include 1,000 targets in the second release of CEFR-LS. Also, judgments by several annotators should

[9] Stanford CoreNLP (Manning et al., 2014) was used for part-of-speech tagging and the results were matched with the wordlist for CEFR levels.

[10] LDC catalog number: LDC2006T13

be added for each candidate. Another possible extension is to include phrases in the dataset (Ganitkevitch et al., 2013; Pavlick and Callison-Burch, 2016). The current dataset is available at our website[11], which will be updated when the 2nd version is ready.

Acknowledgements

This research has been supported by the KDDI Foundation. We thank Professor Christopher G. Haswell for his contribution on annotation as well as valuable comments and discussions with us.

Bibliographical References

Biran, O., Brody, S., and Elhadad, N. (2011). Putting it Simply: A Context-Aware Approach to Lexical Simplification. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–501, June.

Coster, W. and Kauchak, D. (2011a). Learning to Simplify Sentences Using Wikipedia. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–9, June.

Coster, W. and Kauchak, D. (2011b). Simple English Wikipedia: A New Text Simplification Task. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 665–669, June.

Ganitkevitch, J., Durme, B. V., and Callison-Burch, C. (2013). PPDB: The Paraphrase Database. In *Proc. of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 758–764, June.

Glavas, G. and Stajner, S. (2015). Simplifying Lexical Simplification: Do We Need Simplified Corpora? In *Proc. of Annual Meeting of the Association for Computational Linguistics and International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 63–68, July.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a Lexical Simplifier Using Wikipedia. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 458–463, June.

Laufer, B. (1989). What Percentage of Text-Lexis is Essential for Comprehension? In *Chapter 25 in Special language: From humans thinking to thinking machines*, pages 316–323. Multilingual Maters.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–60, June.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. In *Proc. of International Conference on Learning Representations (ICLR)*, May.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, December.

Paetzold, G. H. and Specia, L. (2017). Lexical Simplification with Neural Ranking. In *Proc. of Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 34–40, April.

Pavlick, E. and Callison-Burch, C. (2016). Simple PPDB: A Paraphrase Database for Simplification. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 143–148, August.

Takada, S., Arase, Y., and Uchida, S. (2017). Lexical Simplification for English Education: A Collocation-based Approach. In *Proc. of Annual Meeting of the Association for Natural Language Processing (ANLP)*, pages 939–942, March (in Japanese).

Language Resource References

Cambridge University Press. (2015). *English Vocabulary Profile*. `http://www.englishprofile.org/wordlists`.

Yukio Tono. (2016). *CEFR-J Wordlist Version* 1.3. Tokyo University of Foreign Studies, `http://www.cefr-j.org/download.html`.

[11] `http://www-bigdata.ist.osaka-u.ac.jp/arase/pj/lex-simplification.zip`