Complex Word Identification Based on Frequency in a Learner Corpus

Tomoyuki Kajiwara[†‡]
[†]Institute for Datability Science
Osaka University
Osaka, Japan
kajiwara@ids.osaka-u.ac.jp

Mamoru Komachi[‡]
[‡]Graduate School of Systems Design
Tokyo Metropolitan University
Tokyo, Japan
komachi@tmu.ac.jp

Abstract

We introduce the TMU systems for the complex word identification (CWI) shared task 2018. TMU systems use random forest classifiers and regressors whose features are the number of characters and words and the frequency of target words in various corpora. Our simple systems performed best on 5 of the 12 tracks. Ablation analysis confirmed the usefulness of a learner corpus for a CWI task.

1 Introduction

Lexical simplification (Paetzold and Specia, 2017) is one of the approaches for text simplification (Shardlow, 2014), which facilitates children and language learners' reading comprehension. Lexical simplification comprises the following steps:

1. Complex word identification
2. Substitution generation
3. Substitution selection
4. Substitution ranking

In this study, we work on complex word identification (CWI) (Shardlow, 2013), a subtask of lexical simplification.

Previous studies (Specia et al., 2012; Paetzold and Specia, 2016a) concluded that the most effective way to estimate word difficulty is to count the word frequency in a corpus. However, they counted the word frequency in corpora written by native speakers, such as Wikipedia. Language learners tend to use simple words as compared to native speakers. Therefore, we expect the word frequency in the learner corpus to be a useful feature for CWI.

Our CWI system considers the word frequency in a learner corpus as well as in corpora written by native speakers. We use the Lang-8 corpus[1] (Mizumoto et al., 2011), a learner corpus that can be used on a large-scale in many languages.

2 CWI Shared Task 2018

In CWI shared tasks, systems predict whether words in a given context are complex or non-complex for a non-native speaker. The first CWI shared task (Paetzold and Specia, 2016a; Zampieri et al., 2017) contained only English data designed for non-native English speakers. Totally, 20 annotators were assigned to each instance in the training set. However, in the test set, only one annotator was assigned to each instance. By contrast, the CWI shared task 2018 (Yimam et al., 2018) used a multilingual dataset (Yimam et al., 2017a,b) having all instances annotated by multiple annotators. This shared task was divided into two tasks (binary and probabilistic classification) and the following four tracks:

- English monolingual CWI
- Spanish monolingual CWI
- German monolingual CWI
- Multilingual CWI with a French test set

The English dataset contained a mixture of professionally written news, non-professionally written news (WikiNews), and Wikipedia articles. Datasets for languages excluding English were from Wikipedia articles. Tables 1 and 2 display the dataset and the number of instances, respectively.

[1] http://lang-8.com/

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 195–199
New Orleans, Louisiana, June 5, 2018.

Sentence	Target	Label	Probability
According to Goodyear, a neighbor heard gun shots.	shots	0	0.00
According to Goodyear, a neighbor heard gun shots.	according to	1	0.05
A lieutenant who had defected was also killed in the clashes.	defected	1	0.45
A bad part of the investigation is we may not get the why.	investigation	1	0.95

Table 1: Example instances of the English dataset.

Dataset		Train	Dev	Test
English	(News)	14,002	1,764	2,095
English	(WikiNews)	7,746	870	1,287
English	(Wikipedia)	5,551	694	870
Spanish	(Wikipedia)	13,750	1,622	2,233
German	(Wikipedia)	6,151	795	959
French	(Wikipedia)	0	0	2,251

Table 2: Number of instances.

	Feature
1	Number of characters
2	Number of words
3	Frequency of target in the Wikipedia corpus
4	Frequency of target in the WikiNews corpus
5	Frequency of target in the Lang-8 corpus
6	Probability of target in the Wikipedia corpus
7	Probability of target in the WikiNews corpus
8	Probability of target in the Lang-8 corpus

Table 3: Our features.

	Wikipedia	WikiNews	Lang-8
English	94,872,197	325,038	3,261,441
Spanish	20,197,778	107,289	185,677
German	44,280,830	145,326	160,110
French	26,224,666	135,845	181,004

Table 4: Number of sentences.

2.1 Binary Classification Task

Labels in the binary classification task were assigned as follows:

0: simple word (none of the annotators marked the word as difficult)

1: complex word (at least one annotator marked the word as difficult)

We evaluated the systems using the macro-averaged F1-score.

2.2 Probabilistic Classification Task

Labels in the probabilistic classification task were assigned as the proportion of annotators identifying the target as complex. Systems were evaluated using the MAE (mean absolute error).

3 TMU Systems

According to previous studies (Specia et al., 2012; Paetzold and Specia, 2016a), we estimated the word difficulty by counting word frequency.

3.1 Classifiers

We used random forest classifiers and random forest regressors for binary classification tasks and probabilistic classification tasks, respectively. We examined all combinations of the following hyper-parameters[2]:

- `n_estimators`: $\{10, 50, 100, 500, 1000\}$
- `max_depth`: $\{5, 10, 15, 20, \infty\}$
- `min_samples_leaf`: $\{1, 5, 10, 15, 20\}$

[2]http://scikit-learn.org/

3.2 Features

Table 3 shows all the features used by our systems.

First, we used the heuristics that the longer words are more complex to understand as the first feature. For example, Flesch reading ease (Flesch, 1948), frequently used in research on text simplification, uses this heuristics.

Second, as shown in Table 1, the target includes words and phrases. As long phrases tend to be less frequent, we used the number of words as the second feature.

Others features (3-8) are based on the frequency of targets in a corpus. We counted frequencies from texts written by native speakers and language learners. Language learners are more likely to use simple words than native speakers. Therefore, we expected word frequency in the learner corpus to be a useful feature for CWI. As a text written by native speakers, we counted the frequency from Wikipedia and WikiNews. By contrast, as a text written by language learners, we counted the frequency from the Lang-8 corpus (Mizumoto et al., 2011). The Lang-8 corpus contains texts before and after corrections written by learners and native speakers, respectively. We use the former.

News	Wikipedia	WikiNews	Spanish	German	French
.874 Camb	.812 Camb	.840 Camb	.770 TMU	.745 TMU	.760 CoastalCPH
.864 ITEC	.797 NILC	.831 NLP-CIC	.767 NLP-CIC	.743 SB@GU	.747 TMU
.864 NILC	.792 UnibucKernel	.828 NILC	.764 ITEC	.693 hu-berlin	.627 SB@GU
.863 TMU	.783 SB@GU	.816 CFILT-IITB	.746 CoastalCPH	.662 CoastalCPH	.574 hu-berlin
.855 NLP-CIC	.782 ITEC	.813 UnibucKernel	.728 SB@GU	.555 Gillin Inc.	
.848 CFILT_IITB	.776 CFILT_IITB	.811 ITEC	.708 hu-berlin		
.833 SB@GU	.772 NLP-CIC	.803 SB@GU	.680 Gillin Inc.		
.826 hu-berlin	.762 TMU	.787 TMU			
.824 Gillin Inc.	.745 hu-berlin	.766 hu-berlin			
.818 UnibucKernel	.740 LaSTUS	.749 LaSTUS			
.810 LaSTUS	.721 CoastalCPH	.732 Gillin Inc.			
	.660 Gillin Inc.				

Table 5: Performance on the binary classification task. Systems are ranked by their macro-averaged F1-score.

News	Wikipedia	WikiNews	Spanish	German	French
.051 TMU	.074 Camb	.067 Camb	.072 TMU	.061 TMU	.066 CoastalCPH
.054 ITEC	.081 ITEC	.070 TMU	.073 ITEC	.075 CoastalCPH	.078 TMU
.056 Camb	.082 NILC	.071 ITEC	.079 CoastalCPH	.191 Gillin Inc.	
.059 NILC	.093 TMU	.073 NILC	.251 Gillin Inc.		
.153 SB@GU	.176 SB@GU	.165 SB@GU			
.281 Gillin Inc.	.316 Gillin Inc.	.289 Gillin Inc.			

Table 6: Performance on the probabilistic classification task. Systems are ranked by their MAE score.

3.3 Experimental Settings

The dump data of Wikipedia and WikiNews on December 01, 2017, were downloaded and divided into sentences using WikiExtractor[3] and NLTK[4]. All corpora (Train / Dev / Test and Wikipedia / WikiNews / Lang-8) were tokenized and lowercased in the script of the statistical machine translation tool Moses[5] (Koehn et al., 2007). Table 4 displays the number of sentences in each corpus.

4 Results

Tables 5 and 6 present the official evaluation results. In Table 5, systems are ranked by their macro-averaged F1-score for the binary classification task. TMU systems ranked first in Spanish and German, and second in French. In Table 6, systems are ranked by their MAE score for the probabilistic classification task. TMU systems ranked first in Spanish, German, and English news track and second in English WikiNews track.

4.1 Ablation Analysis of Freq. and Proba.

Frequency and probability are similar features. Table 7 indicates that although the probability features are more important than the frequency features, systems can yield better performance by considering both features.

4.2 Ablation Analysis of Corpora

We examined which corpus provides important features. Table 8 shows the most important features obtained from the Lang-8 corpus. Remarkably, the largest Wikipedia corpus does not contribute significantly to performance.

5 Related Work

Although our systems (random forest with length and frequency of the target word) are simple, they achieve competitive results. In the first CWI shared task 2016, numerous systems (Brooke et al., 2016; Davoodi and Kosseim, 2016; Mukherjee et al., 2016; Zampieri et al., 2016; Ronzano et al., 2016) used random forest classifiers. The length (Wróbel, 2016; Paetzold and Specia, 2016b; Malmasi and Zampieri, 2016; Malmasi et al., 2016; Zampieri et al., 2016; Ronzano et al., 2016; Palakurthi and Mamidi, 2016; Quijada and Medero, 2016; Konkol, 2016) and frequency (Wróbel, 2016; Paetzold and Specia, 2016b; Brooke et al., 2016; Zampieri et al., 2016; Ronzano et al., 2016; Palakurthi and Mamidi, 2016; Quijada and Medero, 2016; Konkol, 2016; Kauchak, 2016) of the target word were the basic

[3]https://github.com/attardi/wikiextractor/
[4]http://www.nltk.org/
[5]https://github.com/moses-smt/mosesdecoder

	News	Wikipedia	WikiNews	Spanish	German	French	Average
Binary Classification Task (macro-averaged F1)							
All Features	0.863	0.762	0.787	0.770	0.745	0.747	0.779
w/o Frequency	0.864	0.770	0.798	0.774	0.742	0.693	0.774
w/o Probability	0.860	0.767	0.803	0.779	0.753	0.663	0.771
Probabilistic Classification Task (MAE)							
All Features	0.051	0.093	0.070	0.072	0.061	0.078	0.071
w/o Frequency	0.052	0.090	0.073	0.071	0.059	0.099	0.074
w/o Probability	0.051	0.094	0.070	0.072	0.061	0.111	0.077

Table 7: Ablation analysis of frequency and probability features.

	News	Wikipedia	WikiNews	Spanish	German	French	Average
Binary Classification Task (macro-averaged F1)							
All Features	0.863	0.762	0.787	0.770	0.745	0.747	0.779
w/o Wikipedia	0.860	0.741	0.790	0.758	0.757	0.748	0.776
w/o WikiNews	0.858	0.750	0.788	0.756	0.748	0.746	0.774
w/o Lang-8	0.859	0.764	0.786	0.743	0.752	0.735	0.773
Probabilistic Classification Task (MAE)							
All Features	0.051	0.093	0.070	0.072	0.061	0.078	0.071
w/o Wikipedia	0.053	0.091	0.072	0.073	0.060	0.079	0.071
w/o WikiNews	0.051	0.092	0.070	0.073	0.061	0.075	0.070
w/o Lang-8	0.052	0.093	0.073	0.075	0.062	0.076	0.072

Table 8: Ablation analysis of corpora.

features of the CWI shared task 2016. These are used as baselines, and a majority of the systems use them as part of their features.

While previous works counted the word frequency in corpora such as Wikipedia, which is written by native speakers, we used corpora written by language learners. As anticipated, the word frequency in the learner corpus proved to be a vital feature in the CWI task.

6 Conclusion

We explained the TMU systems for CWI shared task 2018. Our systems performed best on 5 of the 12 tracks using only simple features.

Previous studies concluded that the most effective way to estimate word difficulty is to count the word frequency in a corpus. However, it was not clear what kind of corpus is useful for counting word frequencies. We discussed the usefulness of a learner corpus for the CWI task for the first time. As anticipated, the word frequency counted from the learner corpus worked better than that from the in-domain corpus written by the native speakers for the CWI task.

Acknowledgements

We would like to thank Xi Yangyang for granting use of extracted texts from Lang-8.

References

Julian Brooke, Alexandra Uitdenbogerd, and Timothy Baldwin. 2016. Melbourne at SemEval 2016 Task 11: Classifying Type-level Word Complexity using Random Forests with Corpus and Word List Features. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 975–981.

Elnaz Davoodi and Leila Kosseim. 2016. CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 982–985.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

David Kauchak. 2016. Pomona at SemEval-2016 Task 11: Predicting Word Complexity Based on Corpus Frequency. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1047–1051.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra

Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180.

Michal Konkol. 2016. UWB at SemEval-2016 Task 11: Exploring Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1038–1041.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 996–1000.

Shervin Malmasi and Marcos Zampieri. 2016. MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifier. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 991–995.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155.

Niloy Mukherjee, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. JU_NLP at SemEval-2016 Task 11: Identifying Complex Words in a Sentence. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 986–990.

Gustavo Paetzold and Lucia Specia. 2016a. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. SV000gg at SemEval-2016 Task 11: Heavy Gauge Complex Word Identification with System Voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 969–974.

Gustavo Paetzold and Lucia Specia. 2017. A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Ashish Palakurthi and Radhika Mamidi. 2016. IIIT at SemEval-2016 Task 11: Complex Word Identification using Nearest Centroid Classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1017–1021.

Maury Quijada and Julie Medero. 2016. HMC at SemEval-2016 Task 11: Identifying Complex Words Using Depth-limited Decision Trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1034–1037.

Francesco Ronzano, Ahmed Abura'ed, Luis Espinosa Anke, and Horacio Saggion. 2016. TALN at SemEval-2016 Task 11: Modelling Complex Words by Contextual, Lexical and Semantic Features. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1011–1016.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *Proceedings of the ACL 2013 Student Research Workshop*, pages 103–109.

Matthew Shardlow. 2014. A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications, Special Issue on Natural Language Processing*, pages 58–70.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proceedings of the *SEM 2012: The First Joint Conference on Lexical and Computational Semantics*, pages 347–355.

Krzysztof Wróbel. 2016. PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 953–957.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017a. CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 401–407.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017b. Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 59–63.

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for ComplexWord Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1001–1005.