1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 14–19
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Incorporating Multiword Expressions in Phrase Complexity Estimation

Sian Gooding,[2] **Shiva Taslimipoor,**[1,2] **Ekaterina Kochmar**[1,2]
[1] ALTA Institute
[2] Department of Computer Science and Technology, University of Cambridge
{shg36, st797, ek358}@cam.ac.uk

Abstract

Multiword expressions (MWEs) were shown to be useful in a number of NLP tasks. However, research on the use of MWEs in lexical complexity assessment and simplification is still an under-explored area. In this paper, we propose a text complexity assessment system for English, which incorporates MWE identification. We show that detecting MWEs using state-of-the-art systems improves predicting complexity on an established lexical complexity dataset.

Keywords: Text simplification, MWE, Lexical simplification

1. Introduction

Complex Word Identification (CWI) is a well-established task in natural language processing, which deals with automated identification of words that a reader might find difficult to understand (Shardlow, 2013). As such, it is often considered the first step in a lexical simplification pipeline. For instance, after a CWI system identifies *sweeping* in:

(1) Prime Minister's government took the *sweeping* action

as complex, a simplification system might suggest replacing it with a simpler alternative, for example with *wide* or *broad*. However, CWI systems so far have been focusing on complexity identification at the level of individual words (Shardlow, 2013; Gooding and Kochmar, 2018; Yimam et al., 2018). At the same time, there is extensive evidence that complexity often pertains to expressions consisting of more than one word. Consider *ballot stuffing* in the following example from the dataset of Yimam et al. (2017):

(2) There have been numerous falsifications and *ballot stuffing*

A CWI system aimed at individual complex word identification would be of a limited use in this case, as trying to simplify *ballot stuffing* on an individual word basis is likely to produce nonsensical or semantically different expressions like *ballot *filling* or *vote stuffing*. *Ballot stuffing* is an example of a *multiword expression* (MWE), which has idiosyncratic interpretation that crosses word boundaries or spaces (Sag et al., 2002). Despite the fact that special consideration of MWEs has been shown to improve results in parsing (Constant et al., 2017), machine translation (Constant et al., 2017; Carpuat and Diab, 2010), keyphrase/index term extraction (Newman and Baldwin, 2012), and sentiment analysis (Williams et al., 2015) and is likely to improve the quality of lexical simplification approaches (Hmida et al., 2018), not much research addressed complexity identification in MWEs (Ozasa et al., 2007; François and Watrin, 2011).

In this paper, we show that identification of MWEs is a crucial step in a lexical simplification pipeline, and in particular it is important at the stage of lexical complexity assessment. In addition, MWEs span a wide range of various expressions, including verbal constructions (*wind down*, *set aside*), nominal compounds (*sledge hammers*, *peace treaty*), named entities (*Barack Obama*, *Los Angeles*), and fixed phrases (*brothers in arms*, *show of force*), among others. Such expressions can be challenging, with various degrees of complexity, for both native and non-native readers. We show that identifying the type of an MWE is helpful at the complexity assessment stage. We also argue that knowing types of MWEs can further assist in selecting an appropriate simplification strategy: for instance, in case of many named entity MWEs and some nominal compounds like *prime minister* the best simplification strategy might consist in providing a reader with a link to a Wikipedia entry.

We present a comprehensive system that:

- discovers MWEs in text;
- identifies MWE type using linguistic patterns; and
- incorporates MWE type into a lexical complexity assessment system.

Our system is trained on a novel lexical complexity dataset for English annotated with the types of MWEs (Kochmar et al., 2020), [1] consisting of 4732 expressions extracted from the complexity-annotated dataset of Yimam et al. (2017). We discuss this dataset in Section 2. Section 3. details our approach to MWE identification. We then present our lexical complexity assessment system in Section 4., and discuss the results of both MWE detection and complexity assessment systems in Section 5.

2. Complex Phrase Identification Dataset

The dataset of Yimam et al. (2017) is the most comprehensive dataset annotated for lexical complexity in context. It consists of 34879 lexemes annotated as simple or complex by 20 annotators, 10 of which are native and other 10 are non-native speakers of English, sourced via Amazon Mechanical Turk. Annotators were presented with text passages of $5-10$ sentences from texts of one of three genres (professionally written NEWS, WIKINEWS written by amateurs,

[1] https://github.com/ekochmar/MWE-CWI

MWE Type	Examples	%
MW compounds:	*life threatening*, *property sector*	26.88
MW named entities:	*Alawite sect*, *Formica Fusca*	10.50
Verb-particle and other phrasal verbs:	*close down*, *get rid of*	2.51
Fixed phrase:	*conflict of interest*, *et al.*	1.52
Semi-fixed VP:	*flexed <their> muscles*, *close <the> deal*	0.82
Verb-preposition:	*morph into*, *shield against*	0.72
PP modifier:	*upon arrival*, *within our reach*	0.70
Conjunction / Connective:	*thus far*, *according to*	0.34
Verb-noun(-preposition):	*provides access to*, *bid farewell*	0.32
Coordinated phrase:	*shock and horror*, *import and export*	0.23
Support verb:	*make clear*, *has taken steps*	0.15
Not MWE:	*vehicle rolled over*, *IP address is blocked*	46.09
Not MWE but contains MWE(s):	*collapsed property sector*, *interior ministry troops*	9.21

Table 1: Classes of MWEs annotated in the dataset of Kochmar et al. (2020)

and WIKIPEDIA articles), and were asked to highlight words and sequences of words up to 50 characters in length that they considered difficult to understand. As a result, Yimam et al. (2017) collected a dataset of 30147 individual words and 4732 "phrases" annotated as simple or complex in context. The annotation follows one of the two settings: under *binary* setting a lexeme receives a label of 1 even if a single annotator selected it as complex (0 if none of the annotators considered it complex), and under *probabilistic* setting a lexeme receives a label on the scale of $[0.0, 0.05, ..., 1.0]$ representing the proportion of annotators among 20 that selected an item as complex.

During annotation, annotators were allowed to select any sequence of words, which resulted in selection of expressions that do not form MWEs proper (for instance, *his drive*), as well as sentence fragments and sequences of unrelated words (for instance, *authorities should annul the*). Since the annotators in Yimam et al. (2017) were not instructed to select proper MWEs in this data, Kochmar et al. (2020) first re-annotated the selection of 4732 sequences longer than one word from the original dataset with their MWE status and type.

In this annotation experiment, Kochmar et al. (2020) followed the annotation instructions and distinguished between the MWE types from Schneider et al. (2014), with a few modifications:

- Additional types for "phrases" that are not MWE proper were introduced. These types include `Not MWE` for cases like *authorities should annul the*, and `Not MWE but contains MWE(s)` for longer non-MWE expressions that contain MWEs as sub-units: for example, *collapsed property sector*.
- Two categories, `verb-particle` and `other phrasal verb`, were merged into one due to lack of distinguishing power between the two from the simplification point of view.
- Categories `phatic` and `proverb` were not used because examples of these types do not occur in this data.

Table 1 presents the full account of MWE types with examples and their distribution in the dataset of Kochmar et al. (2020). The dataset was annotated by 3 annotators, all trained in linguistics, over a series of rounds. The annotators achieved observed agreement of at least 0.70 and Fleiss κ (Fleiss, 1981) of at least 0.7145 across the annotation rounds, which suggests substantial agreement. We refer the readers to the original publication (Kochmar et al., 2020) for more details on the annotation procedure.

3. Multiword Expression Identification

We first need to train an MWE identification system to detect the expressions of interest for our study. MWE identification is the task of discriminating, in context, and linking those tokens that together develop a special meaning. This can be modestly modelled using sequence tagging systems. We experiment with two systems: one is BERT-based transformer (Devlin et al., 2018) for token classification, and the other is the publicly available graph convolutional neural network (GCN) based system, which is reported to achieve state-of-the-art results on MWE identification (Rohanian et al., 2019).

The BERT-based token classification system is designed by adding a linear classification layer on top of the hidden-states output of the BERT architecture. We use the pre-trained model of `bert-base` provided by 'Hugging Face' developers [2] and fine-tune the weights of the whole architecture for a few iterations (i.e. 5 epochs). We use the same configurations that they use for named entity recognition. Among various systems designed to tag corpora for MWEs (Ramisch et al., 2018) the best systems incorporate dependency parse information (Al Saied et al., 2017; Rohanian et al., 2019). The GCN-based system that we employ consists of GCN and LSTM layers with a linear classification layer on top. As in the original system, we use ELMo for input representation.

Since our complexity estimation dataset is not originally designed for MWE identification, we augment our training data with the STREUSLE dataset which is comprehensively annotated for MWEs (Schneider and Smith, 2015). In Section 5. we show how this addition helps better identification of MWEs.

[2] `https://github.com/huggingface/transformers`

Once MWEs are identified in text, their types are predicted based on linguistic patterns. For instance, an MWE detection system identifies *woke up* as an MWE in *He woke up in the morning as usual*. A linguistic patterns-based system then uses the information about the parts-of-speech in this expression to predict its type as `verb-particle and other phrasal verbs`. Next, the predicted MWE together with its type is passed on to the lexical complexity assessment system that assesses the complexity of the expression (see Section 4.).

In Section 5. we first compare the results of the two MWE identification systems. Then we use the best one in evaluating the performance of complexity assessment.

4. MWE Complexity Assessment Systems

We build a baseline MWE complexity system, whose goal is to assign a complexity score to identified MWEs. The complexity assessment system is trained on phrases that have been annotated as MWEs in our dataset, and tested using the MWEs extracted from the test portion of the shared task dataset (Yimam et al., 2018).

We run experiments using the probabilistic labels, which represent the complexity of phrases on a scale of $[0.0...0.70]$,[3] representing the proportion of 20 annotators that found a phrase complex. The MWE complexity assessment system is a supervised feature-based model.

4.1. Features

Our complexity assessment system relies on 6 features. First, we include two traditional features found to correlate highly with word complexity in previous research: *length* and *frequency*. These are adapted for phrases by considering (1) the number of words instead of the number of characters for *length*, and (2) using the average frequency of bigrams within the phrase, which is calculated using the Corpus of Contemporary American English (Davies, 2009) for *frequency*. Average bigram frequency is used rather than n-gram frequency to account for the differences in MWE lengths and to increase feature coverage.

The second category of features focuses on the complexity of words contained within the MWE. We use an open source system of Gooding and Kochmar (2019) to tag words with a complexity score. Since this system does not directly assign complexity scores to MWEs, we use the highest word complexity within the phrase as well as the average word complexity as features.

The source genre of the sentence where a phrase occurs (NEWS, WIKINEWS or WIKIPEDIA) is used as another feature, as we hypothesise that different domains (e.g., more general for the NEWS vs. more technical for the WIKIPEDIA articles) may challenge readers to a different extent. Finally, following Kochmar et al. (2020), who show that different types of MWEs show different complexity levels, we use the type of MWE predicted by the linguistic patterns-based system as a feature. An example of the feature set for the phrase *sledge hammers* is shown in Table 2.

[3]The upper bound on this scale reflects the fact that at most 14 annotators agreed that a particular phrase is complex.

	sledge hammers
MWE	MW Compounds
Length	2
Freq	39
Max CW	0.70
Mean CW	0.60
Genre	News

Table 2: Complexity prediction feature set for *sledge hammers*

4.2. System Implementation

A set of standard regression algorithms from the `scikit-learn`[4] library are applied to the dataset. Model predictions are rounded to the closest 0.05 interval. The best performing model, identified via stratified 5-fold cross validation, uses a Multi-layer Perceptron regressor with 6 hidden layers and the `lbfgs` optimiser, used due to the size of the dataset.

5. Experiments

5.1. MWE Identification Results

We report the results of our MWE identification systems compared to the gold standard annotation which is explained in Section 2. We evaluate the systems in terms of the MWE-based precision, recall and F1-score which are defined in Savary et al. (2017). MWE-based evaluation measures count the strict matching between the prediction and the gold labels where every component of an MWE should be correctly tagged in order for it to be considered true positive. In Table 3, we report the MWE-based measures for both positive (MWE) and negative (non-MWE) classes.[5]

As can be seen in Table 3, the graph convolutional neural network-based (GCN) system outperforms Bert-transformer token classification for identifying MWEs. We can also see that the addition of external MWE-annotated data from STREUSLE helps improving the overall results. As expected, the data augmentation is especially effective in increasing recall as well as the overall F-measure.

The best-performing system, GCN trained on both our MWE data and STREUSLE dataset, achieves the highest F1-scores of 0.72 on `not MWE` and 0.60 on `MW compounds` classes, which are also the most prevalent in our data. At the same time, it finds detection of less frequent classes like `verb-preposition`, `verb-noun(-preposition)` and `conjunction/connective` more challenging.

5.2. End-to-end Complexity System Results

We use a pipeline system consisting of three stages: (1) *MWE identification*, (2) *MWE type prediction*, and (3) *MWE complexity prediction*. In Table 4 we report the results on the MWE proportion of the 2018 shared task test sets (Yimam et

[4]`https://scikit-learn.org`

[5]The negative class (non-MWEs) includes expressions (sequences of words) that are present in the dataset of Yimam et al. (2018) but are not tagged as MWEs in Kochmar et al. (2020), e.g. *authorities should annul the*.

training data	model	MWE class			non-MWE class		
		P	R	F1	P	R	F1
Our data train	GCN	93.67	37.37	53.43	66.03	97.97	78.89
	BERT-transformer	90.62	29.29	44.27	63.16	97.56	76.68
Our data train + STREUSLE	GCN	90.80	39.90	**55.44**	66.67	96.75	**78.94**
	BERT-transformer	95.95	35.86	52.21	65.68	98.78	78.90

Table 3: Performance of MWE identification systems in the development phase

Test Set	MAE	
	System	CAMB
(3) Complexity Prediction		
News (133)	**0.0688**	0.0767
Wikipedia (84)	**0.0671**	0.0734
WikiNews (79)	0.0375	**0.0327**
(2,3) MWE Type Prediction + Complexity Prediction		
News (133)	**0.0745**	0.0767
Wikipedia (84)	**0.0720**	0.0734
Wikinews (79)	0.0474	**0.0327**
(1,2,3) MWE Identification + MWE Type Prediction + Complexity Prediction		
News (61)	**0.0889**	0.0984
Wikipedia (27)	**0.1221**	0.1283
WikiNews (23)	**0.0572**	0.0595

Table 4: Complexity assessment system results

al., 2018) for each stage of the pipeline. We compare our results to the strategy used by the winning shared task system CAMB (Gooding and Kochmar, 2018), where all phrases are simply assigned the complexity value of 0.05. This baseline is highly competitive, as 1074 of the 2551 examples have a probabilistic score of 0.05, with 61% of MWEs having a value of 0.00 or 0.05. We use Mean Absolute Error (MAE) as our evaluation metric, following the 2018 Shared Task official evaluation strategy (Yimam et al., 2018). This metric estimates average absolute difference between pairs of the predicted and the gold-standard complexity scores.

The initial results in Table 4 consider *complexity prediction* in isolation, by testing on valid MWEs and providing the gold labels for the MWE types. Our system achieves lower absolute error than the baseline on both NEWS and WIKIPEDIA test sets, but not on the WIKINEWS test set. However, the distribution of probabilistic scores in the WIKINEWS test set is highly skewed, with 79% having scores of 0.05 or 0.00 and the highest complexity score in the dataset being only 0.35; a graph in Figure 1 illustrates the distribution of labels across test sets.

In practice we do not have gold standard labels for the MWE types, therefore we use linguistic pattern analysis to predict the MWE labels. The results of combining type and complexity prediction (2,3) follow the same trend as complexity prediction alone, however they also show a decrease in performance across test sets. As Kochmar et al. (2020) show, the type of MWE is highly informative when considering phrase complexity, therefore misclassification at this stage negatively impacts subsequent complexity prediction. We note that our MWE-type detection system achieves the F1-scores around 0.70 on the `MW named entities`, `PP modifier` and `verb-particle or other phrasal verb` classes, followed by F1-scores around 0.60 for the `MW compounds` and `verb-preposition` classes. The classes that our system most struggles in identifying include `conjunction/connective`, `coordinated phrase` and `verb-noun(-preposition)`.

Finally, we consider the entire pipeline including the initial step (1) of *MWE identification*. As complexity prediction can only be performed on MWEs identified by our system, the size of the test set is reduced, therefore results are not directly comparable to previous stages. However, we note that our system outperforms the baseline across all genres. The baseline performs worse on the MWEs identified by our system as the probabilistic average is higher (0.14 compared to 0.09). A point of interest is that of the MWEs identified by the system, only 0.08% have a complexity value of 0 compared to 18% of the initial test sets. This suggests that the MWE identification step is identifying 'strong' MWEs that are more likely to be considered complex by annotators. This further supports our hypothesis that an MWE identification system can be combined with complexity features into a unified system to provide better complexity identification at the level of phrases.

6. Conclusions

In this paper, we propose a complexity assessment system for predicting complexity of MWEs rather than single word units. We show that augmenting the system with the information about type of expressions improves the performance. Research on lexical complexity assessment would highly benefit from the proposed data and system.

Acknowledgements

The second and third authors' research is supported by Cambridge Assessment, University of Cambridge, via the ALTA

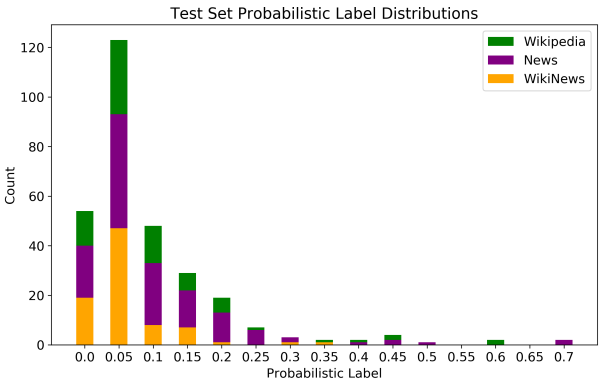

Figure 1: Probabilistic label counts across test sets

Institute. We are grateful to the anonymous reviewers for their valuable feedback.

7. Bibliographical References

Al Saied, H., Candito, M., and Constant, M. (2017). The ATILF-LLF System for Parseme Shared Task: a Transition-based Verbal Multiword Expression Tagger. In *13th Workshop on Multiword Expressions (MWE 2017)*, pages 127–132.

Carpuat, M. and Diab, M. (2010). Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proceedings of NAACL-HLT*, pages 242–245.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: John Wiley, 2nd edition.

François, T. and Watrin, P. (2011). On the contribution of MWE-based features to a readability formula for French as a foreign language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 441–447, Hissar, Bulgaria, September. Association for Computational Linguistics.

Gooding, S. and Kochmar, E. (2018). CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana, June. Association for Computational Linguistics.

Gooding, S. and Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy, July. Association for Computational Linguistics.

Hmida, F., Billami, M., François, T., and Gala, N. (2018). Assisted lexical simplification for french native children with reading difficulties. In *Proceedings of the Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation*.

Kochmar, E., Gooding, S., and Shardlow, M. (2020). Detecting multiword expression type helps lexical complexity assessment. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*.

Newman, David, K. N. L. J. H. and Baldwin, T. (2012). Bayesian text segmentation for index term identification and keyphrase extraction. In *Proceedings of COLING 2012*, pages 2077–2092.

Ozasa, T., Weir, G., and Fukui, M. (2007). Measuring readability for Japanese learners of English. In *Proceedings of the 12th Conference of Pan-Pacific Association of Applied Linguistics*, pages 122–125.

Ramisch, C., Cordeiro, S., Savary, A., Vincze, V., Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., et al. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions.

Rohanian, O., Taslimipoor, S., Kouchaki, S., Ha, L. A., and Mitkov, R. (2019). Bridging the gap: Attending to discontinuity in identification of multiword expressions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2692–2698.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. *Lecture Notes in Computer Science*, 2276:1–15.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., Qasemizadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., et al. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47.

Schneider, N. and Smith, N. A. (2015). A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1537–1547.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., and Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461. European Language Resources Association (ELRA), May.

Shardlow, M. (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, August. Association for Computational Linguistics.

Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., and Spasić, I. (2015). The role of idioms in sentiment analysis. *Expert Systems with Applications*, 42(21):7375–7385.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). 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 (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*,

pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.