Complex Word Identification as a Sequence Labelling Task

Sian Gooding
Dept of Computer Science and Technology
University of Cambridge
shg36@cam.ac.uk

Ekaterina Kochmar
ALTA Institute
University of Cambridge
ek358@cam.ac.uk

Abstract

Complex Word Identification (CWI) is concerned with detection of words in need of simplification and is a crucial first step in a simplification pipeline. It has been shown that reliable CWI systems considerably improve text simplification. However, most CWI systems to date address the task on a word-by-word basis, not taking the context into account. In this paper, we present a novel approach to CWI based on sequence modelling. Our system is capable of performing CWI in context, does not require extensive feature engineering and outperforms state-of-the-art systems on this task.

1 Introduction

Lexical complexity is one of the main aspects contributing to overall text complexity (Dubay, 2004). It is typically addressed with lexical simplification (LS) systems that aim to paraphrase and substitute complex terms for simpler alternatives. Previous research has shown that Complex Word Identification (CWI) considerably improves lexical simplification (Shardlow, 2014; Paetzold and Specia, 2016a). This is achieved by identifying complex terms in text prior to word substitution. The performance of a CWI component is crucial, as low recall of this component might result in an overly difficult text with many missed complex words, while low precision might result in meaning distortions with an LS system trying to unnecessarily simplify non-complex words (Shardlow, 2013).

CWI has recently attracted attention as a standalone application, with at least two shared tasks focusing on it. Current approaches to CWI, including state-of-the-art systems, have a number of limitations. First of all, CWI systems typically address this task on a word-by-word basis, using a large number of features to capture the complexity of a word. For instance, the CWI system by Paetzold and Specia (2016c) uses a total of 69 features, while the one by Gooding and Kochmar (2018) uses 27 features. Secondly, systems performing CWI in a static manner are unable to take the context into account, thus failing to predict word complexity for polysemous words as well as words in various metaphorical or novel contexts. For instance, consider the following two contexts of the word *molar* from the CWI 2018 shared task (Yimam et al., 2018). *Molar* has been annotated as complex in the first context (resulting in the binary annotation of 1) by 17 out of 20 annotators (thus, the “probabilistic” label of 0.85), and as non-complex (label 0) in the second context:

Contexts	Bin	Prob
Elephants have four **molars**...	1	0.85
... new **molars** emerge in the back of the mouth.	0	0.00

The annotators may have found the second context simpler on the whole, as *molars* is surrounded by familiar words that imply the meaning (e.g., *mouth*), whereas *elephants* is a rarer and less semantically similar co-occurrence. Such context-related effects are hard to capture with a CWI system that only takes word-level features into account. Thirdly, CWI systems that only look at individual words cannot grasp complexity above the word level, for example, when a whole phrase is considered complex.

In this paper, we apply a novel approach to the CWI, based on sequence labelling.[1] We show that our system is capable of:

- taking word context into account;
- relying on word embeddings only, thus eliminating the need for extensive feature engineering;
- detecting both complex words and phrases;

[1] Trained models are available at: https://github.com/siangooding/cwi

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1148–1153
Florence, Italy, July 28 - August 2, 2019.

- not requiring genre-specific training and representing a one-model-fits-all approach.

2 Related Work

2.1 Complex Word Identification

Early studies on CWI address this task by either attempting to simplify all words (Thomas and Anderson, 2012; Bott et al., 2012) or setting a frequency-based threshold (Zeng et al., 2005; Elhadad, 2006; Biran et al., 2011). Horn et al. (2014) show that the former approach may miss up to one third of complex words due to its inability to find simpler alternatives, and Shardlow (2013) argues that a simplify-all approach might result in meaning distortions, but the more resource-intensive threshold-based approach does not necessarily perform significantly better either. At the same time, Shardlow (2013) shows that a classification-based approach to CWI is the most promising one. Most of the teams participating in the recent CWI shared tasks also use classification approaches with extensive feature engineering.

The first shared task on CWI at SemEval 2016 (Paetzold and Specia, 2016b) used data from several simplification datasets, annotated by non-native speakers. In this data, about 3% of word types and 11% of word tokens, if contexts are taken into account, are annotated as complex (Paetzold and Specia, 2016b). The CWI 2018 shared task (Yimam et al., 2018) used the data from Wikipedia, news sources and unprofessionally written news, derived from the dataset of Yimam et al. (2017). The dataset was annotated by 10 native and 10 non-native speakers, and, depending on the source of the data, contains 40% to 50% words labelled as complex in context. The dataset contains words and phrases with two labels each. The first label represents binary judgement with *bin*=1 if at least 1 annotator marked the word as complex in context, and *bin*=0 otherwise. The second label is a "probabilistic" label representing the proportion of the 20 annotators that labelled the item as complex. The importance of context when considering word complexity is exemplified well in this dataset, as 11.34% of items have different binary labels depending on the context they are used in. When considering probabilistic annotations, of the items labelled in different contexts 10.96% have at least a 5-annotator difference in complexity score in differing contexts. The dataset contains 104 instances with a 10-annotator difference between scores based on the context of the word. For instance, *suspicion* has been annotated 23 times:

Word	Unique	Max	Min	σ
suspicion	16	0.95	0.15	0.25

Of the 23 probabilistic annotations for *suspicion* 70% are unique. *Max* and *min* values show the largest difference in annotations for this word in context, with 19 annotators labelling it complex in one scenario and only 3 in another. Finally, σ represents the standard deviation of the probabilistic annotations for this word.

In this paper, we use the data from the CWI 2018 shared task, which contains annotation for both words and word sequences (called *phrases* in the task), and represents three different genres of text. We focus on the binary setting (complex vs. non-complex) and compare our results to the winning system by Gooding and Kochmar (2018).

2.2 Sequence Labelling

Sequence labelling has been applied successfully to a number of NLP tasks that rely on contextual information, such as named entity recognition, part-of-speech tagging and shallow parsing. Within this framework, the model receives as input a sequence of tokens $(w_1, ..., w_T)$ and predicts a label for each token as output. Typically, the input tokens are first mapped to a distributed vector space, resulting in a sequence of word embeddings $(x_1, ..., x_T)$. The use of word embeddings allows sequence models to learn similar representations for semantically or functionally similar words. Recent advances to sequential model frameworks have resulted in the models' ability to infer representations for previously unseen words and to share information about morpheme-level regularities (Rei et al., 2016).

Sequence labelling models benefit from the use of long short-term memory (LSTM) units (Gers et al., 2000), as these units can capture the long-term contextual dependencies in natural language. A variation of the traditional architecture, bi-directional LSTMs (BiLSTM) (Hochreiter and Schmidhuber, 1997), has proved highly successful at language tasks, as it is able to consider both the left and right contexts of a word, thus increasing the amount of relevant information available to the network. Similarly, the use of secondary learning objectives can increase the number of salient features and access to relevant information. For example, Rei (2017) shows that training a model to jointly predict surrounding words incentivises the discovery of useful features and associations that are unlikely to be discovered otherwise.

From the perspective of CWI, it is clear that context greatly impacts the perceived difficulty of text. In this paper we investigate whether CWI can be framed as a sequence labelling task.

3 Implementation

For our experiments, we use the English part of the CWI datasets from Yimam et al. (2017), which contains texts on professionally written NEWS, amateurishly written WIKINEWS, and WIKIPEDIA articles. The original data includes the annotation for a selected set of content words, which is provided alongside the full sentence and the word span. The annotation contains both binary (*bin*) and "probabilistic" (*prob*) labels as detailed in Section 2:

Sentence	Word	Bin	Prob
They drastically...	drastically	1	0.5

As the sequential model expects the complete word context as an input, we adapt the original format by tokenizing the sentences and including the annotation for *each* word token, using C for the annotated complex words and phrases, and N for those that are either annotated as non-complex in the original data or not included in it (e.g., function words), which results in the following format:

They	N
drastically	C
...	

We opted to use a sequential architecture by Rei (2017), as it has achieved state-of-the-art results on a number of NLP tasks, including error detection, which is similar to CWI in that it identifies relatively rare sequences of words in context. The design of this architecture is highly suited to the task of CWI as: (1) the use of a BiLSTM provides contextual information from both the left and right context of a target word; (2) the context is combined with both word and character-level representations (Rei et al., 2016); (3) this architecture uses a language modelling objective, which enables the model to learn better composition functions and to predict the probability of individual words in context. As previous work on CWI has consistently found word frequency and length to be highly informative features, we choose an architecture which utilises sub-word information and a language modelling objective.

We use 300-dimensional GloVe embeddings as word representations (Pennington et al., 2014) and train the model on randomly shuffled texts from all three genres for 20 iterations. We train the model using word annotations and predict binary word scores using the output label probabilities. If the probability of a word belonging to the complex class is above 0.50, it is considered a complex word. For phrase-level binary prediction, we consider the phrases contained within the dataset. The complex class probability for each word, aside from stop words, is predicted and combined into a final average score. If this average is above a predefined threshold of 0.50 then the phrase is considered complex.

4 Results & Discussion

Results: We report the results obtained with the sequence labelling (SEQ) model for the binary task and compare them to the current state-of-the-art in complex word identification, CAMB system by Gooding and Kochmar (2018), which achieved the best results across all binary and two probabilistic tracks in the CWI 2018 shared task (Yimam et al., 2018). The evaluation metric reported is the macro-averaged F1, as was used in the 2018 CWI shared task (Yimam et al., 2018). For the binary task, both words and phrases are considered correct if the system outputs the correct binary label.

The CAMB system considers words irrespective of their context and relies on 27 features of various types, encoding lexical, syntactic, frequency-based and other types of information about individual words. The system uses Random Forests and AdaBoost for classification, but as Gooding and Kochmar (2018) report, the choice of the features, algorithm and training data depends on the genre. In addition, phrase classification is performed using a 'greedy' approach and simply labelling all phrases as complex.

The results presented in Table 1 show that the SEQ system outperforms the CAMB system on all three genres on the task of binary complex word identification. The largest performance increase for words is on the WIKIPEDIA test set (+3.60%).

Table 1 also shows that on the combined set of words and phrases (*words+phrases*) the two

Test Set	Macro F-Score	
	CAMB	SEQ
Words Only		
NEWS	0.8633	0.8763 (+1.30)
WIKINEWS	0.8317	0.8540 (+2.23)
WIKIPEDIA	0.7780	0.8140 (+3.60)
Words+Phrases		
NEWS	0.8736	0.8763 (+0.27)
WIKINEWS	0.8400	0.8505 (+1.05)
WIKIPEDIA	0.8115	0.8158 (+0.43)

Table 1: SEQ vs. CAMB system results on words only and on words and phrases

systems achieve similar results: the SEQ model beats the CAMB model only marginally, with the largest difference of $+1.05\%$ on the WIKINEWS data. However, it is worth highlighting that the CAMB system does not perform any phrase classification per se and simply marks all phrases as complex. Using the dataset statistics, we estimate that CAMB system achieves precision of 0.64. The SEQ model outperforms the CAMB system, achieving precision of 0.71.

We note that the SEQ model is not only able to outperform the CAMB system on all datasets for both *words only* and *words+phrases*, but it also has a clear practical advantage: the only input information it uses at run time are word embeddings, whereas the CAMB system requires 27 features based on a variety of sources. In addition, the CAMB system needs to rely on individually tailored systems to maximize the results across datasets, whereas the SEQ model is a 'one size fits all' model that is able to work out-of-the-box across all datasets, achieving state-of-the-art performance by harnessing the power of word context, embeddings and character-level morphology.

We additionally compare our results to the recent work by Maddela and Xu (2018), who show an improvement on the CWI systems with the use of additional 'human-based' features. Using an English lexicon of $15,000$ words with word-complexity ratings by human annotators, they are able to improve the scores of the winning CWI system from the 2016 shared task by Paetzold and Specia (2016c), and the nearest centroid (NC) approach by Yimam et al. (2017). They report the best F-score of 74.8 on the combined CWI 2018 shared task testset, achieved using the NC approach augmented with the complexity lexicon. We note that both CAMB and our SEQ model achieve significantly higher results.

Discussion: To further analyze the results achieved by CAMB and SEQ on the test sets, we apply the McNemar statistical test (McNemar, 1947), which is comparable to the widely used paired t-test, and is most suitable for dichotomous dependent variables. Table 2 presents the contingency table for *words only*, and Table 3 for *words+phrases*:

	CAMB Correct	CAMB Wrong
SEQ Correct	a=3002	b=**205**
SEQ Wrong	c=**145**	d=349

Table 2: Contingency table for *words only*

	CAMB Correct	CAMB Wrong
SEQ Correct	a=3443	b=**207**
SEQ Wrong	c=**145**	d=457

Table 3: Contingency table for *words+phrases*

Using the above values, the continuity corrected McNemar test (Edwards, 1948) estimates χ^2 as:

$$\chi^2 = \frac{(|b-c|-1)^2}{(b+c)} \quad (1)$$

According to the test, the SEQ system achieves significantly better results than the CAMB system on *words only* ($p = 0.0016$, $\chi^2 = 9.95$) as well as on *words+phrases* ($p = 0.0011$, $\chi^2 = 10.57$).

349 word tokens, with 289 word types, are incorrectly labelled by both systems (see Table 2). Of these, 166 words are incorrectly identified as complex, and 183 are incorrectly identified as simple. Of the words that are not identified as complex by the SEQ model, 74% are marked as complex by only one annotator out of twenty, and 93% by one or two annotators. This highlights the idiosyncratic nature of the task and why it may be particularly challenging to address the complexity needs of all individuals with a single system.

There are 205 word instances that are correctly classified by the SEQ model, but not by the CAMB system. 34% of these words the CAMB system correctly classifies in other contexts, but not when the context changes, for instance when the same words are used in unusual or metaphorical contexts. Table 4 presents some examples of the contexts where the SEQ model correctly identifies the complexity of the word, but CAMB model fails (LABEL stands for the gold standard label).

Contexts	CAMB	SEQ	LABEL
Successive **waves** of bank sector reforms have failed	0	1	1
Diffraction occurs with all **waves**	0	0	0

Table 4: Context dependent annotations of the word *waves*

We note that the SEQ model is able to correctly identify the complexity of the word *waves* when used in different contexts. The system outputs a score of 0.5692 for the first context (*Successive waves of bank sector [...]*) and 0.4704 for the second (*Diffraction occurs with all waves*), reflecting that the complexity level is dependent on the context.

5 Conclusions

In this paper, we address the limitations of the existing CWI systems. Our SEQ model relies on sequence labelling and outperforms state-of-the-art systems with a one-model-fits-all approach. It is able to take context into account and classify both words and phrases in a unified framework, without the need for expensive feature engineering. Our future research will focus on the relative nature of complexity judgements and will use the SEQ model to predict complexity on a scale. We will also investigate whether the SEQ model may benefit from sources of information other than word embeddings and character-level morphology. Finally, we plan to investigate alternative methods to modelling phrase and multi-word expression complexity.

Acknowledgments

We thank Cambridge English for supporting this research via the ALTA Institute. We are also grateful to the anonymous reviewers for their valuable feedback.

References

Or Biran, Samuel Brody, and Noemie Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: short papers*, pages 496–501.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING 2012: Technical Papers*, pages 357–374.

William H. Dubay. 2004. The Principles of Readability. *Costa Mesa, CA: Impact Information*.

Allen L. Edwards. 1948. Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika*, 13(3):185–187.

Noemie Elhadad. 2006. Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms. In *AMIA Annual Symposium Proceedings*, pages 239–243.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. 2000. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471.

Sian Gooding and Ekaterina Kochmar. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3749–3760.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Gustavo Paetzold and Lucia Specia. 2016a. PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification. In *Proceedings of the first international workshop on Quality Assessment for Text Simplification (QATS)*, pages 1–9. European Language Resources Association (ELRA).

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

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2121–2130.

Marek Rei, Gamal K.O. Crichton, and Sampo Pyysalo. 2016. Attending to characters in neural sequence labeling models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 309–318.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 103–109.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *In Proceedings of the 9th LREC*, pages 1583–1590.

S. Rebecca Thomas and Sven Anderson. 2012. WordNet-Based Lexical Simplification of a Document. In *Proceedings of KONVENS 2012 (Main track: oral presentations)*.

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*, pages 66–78.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 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.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. *Biological and Medical Data Analysis. ISBMDA 2005. Lecture Notes in Computer Science*, volume 3745 of *ISBMDA 2005*, chapter A Text Corpora-Based Estimation of the Familiarity of Health Terminology. Springer, Berlin, Heidelberg.