

What Makes a Concept Complex? Measuring Conceptual Complexity as a Precursor for Text Simplification

Anne Eschenbrücher^[0000–0001–7024–6100]

Research Group in Computational Linguistics
University of Wolverhampton, UK
a.eschenbruecher@wlv.ac.uk

Abstract. Advancements within the field of Text Simplification (TS) have primarily been within syntactic or lexical simplification. However, conceptual simplification has previously been identified as another field of TS that has the potential to significantly improve reading comprehension. The first step to measuring conceptual simplification is the classification of concepts as either complex or simple. This research-in-progress paper proposes a new definition of conceptual complexity alongside a simple machine-learning approach that performs a binary classification task to distinguish between simple and complex concepts. It is proposed that this be a first step when developing new text simplification models that operate on a conceptual level.

Keywords: Text Complexity · Text Simplification · NLP in Educational Applications.

1 Introduction

Text is a constant companion in our daily lives. Without the ability to write down and record knowledge, the world and its people could arguably never have evolved to its current state. Despite the big impact of text on our daily lives, around one in six adults in the United Kingdom has poor literacy skills [1]. In order to combat this problem, texts have long been simplified to improve understanding for all readers [2]. With the advancements in the field of Natural Language Processing (NLP), or more specifically, Text Simplification (TS), true progress has been made especially within the subfields of lexical and syntactic simplifications [2–5]. A text can, however, be difficult to understand not only lexically or syntactically, but also conceptually. As of yet, conceptual text simplification has been paid significantly less attention than its two related subfields, although it has the potential to significantly improve reading comprehension. Conceptual complexity is in itself a concept hard to grasp. Definitions vary from “the importance of distinguishing between ideas in a text and fact burden of a text” [6] to the amount of background knowledge a reader needs to in order to understand the concepts mentioned in the text and how they relate to each other

[7, 8]. Yet another definition refers to a “need for algorithms that can understand what a user needs to know before they can understand a second concept” [9]. According to all these definitions, the amount of background knowledge is what essentially makes a text conceptually complex. However, they all fail to take into account reading processes that occur in the brain.

While reading a text, a reader constantly has to chunk concepts together into larger order units in order to free up working memory space to avoid displacement [10]. In order for this to happen, all concepts and their relations that are involved need to be present simultaneously in short-term memory (STM) [11]. This is where differences between skilled and poor readers become obvious as skilled readers are able to suppress irrelevant information more easily and only store what is truly important, while also being generally better at chunking the relevant information which enables them to comprehend a text more fully than poor readers would [12].

When reading a text, ideally a reader would chunk as much as possible to avoid displacement as this would mean that important concepts are lost from STM [10]. It can therefore be argued that a text is conceptually complex if the working memory’s capacity is frequently at its limit. We assume that closely related concepts are probably easier to chunk together, which is why relatedness should also play an important role. Hence, our working definition of conceptual complexity is as follows: Conceptual complexity is the complexity of a concept by itself, referring to level of abstractness and frequency of use in everyday language, as well as the complexity of a concept’s context, meaning how complex its relations are to other concepts in its direct vicinity.

2 Related Work

One particular challenge within the field of conceptual TS is the estimation of conceptual text complexity. While many studies focus on automatic assessment of lexical and syntactic complexity [13–15], automatic approaches for measuring conceptual complexity are rare. One of the first automated tools that took conceptual complexity into account was the Coh-Metrix tool which uses latent semantic analysis semantic features and coreference indices alongside measures of lexical and syntactic complexity [16]. More recently, the estimation of conceptual complexity has been attempted using knowledge graphs [7], as well as a mix of surface-based and shallow semantic features [8].

Recently, the focus of text simplification systems has shifted from simplifying a text to its easiest possible version to simplifications that are tailored to a specific reader’s needs [5]. So far, text simplification systems that take the concept level into account use mainly extractive approaches that eliminate irrelevant information [17, 18]. Some concepts, however, might be essential to a text and cannot be eliminated which is why an abstractive approach needs to be used that generates new and simpler output. To identify which concepts are complex and need to be replaced by simpler substitutes, it is not sufficient to estimate the complexity of an entire text. Rather, each individual concept needs to

be assigned a complexity score that takes into account the concept’s individual complexity, as well as its context complexity.

3 Methodology

The proposed methodology contains several features that take into account a concept’s relations with other concepts that occur in the same sentence, as well as how complex the concept is by itself. Those features, as well as the model and annotation process are briefly described below.

3.1 Pairwise Features

Features that take into account a concept’s relations to other concepts are computed using WordNet and ConceptNet databases [19, 20]. In order to determine how complex a concept is in relation to other concepts it is being mentioned with, we compute two types of pairwise features.

Firstly, we compute the WordNet path similarity between a concept and all other concepts in a sentence. The mean for each concept is taken as that concept’s score for the respectable context. Secondly, ConceptNet relatedness is computed using the ConceptNet API. Here, we also take the mean for each concept’s score together with all other concepts in that sentence as feature.

In the sentence “Such were Elizabeth Elliot’s sentiments and sensations; such the cares to alloy, the agitations to vary”, we first extract and lemmatise the concepts “sentiment”, “sensations”, “cares” and “agitations”. We then take the first concept “sentiment” and compute e.g. its path similarity with all other concepts in this sentence, finally taking the mean of all those as a feature. We repeat this process for each concept in a given sentence before moving on to the next. The same is repeated to compute ConceptNet relatedness.

3.2 Single Features

A total of eight features is computed for each single concept. Word length is taken into account as more complex meanings tend to be encoded in longer structures [21]. Furthermore, type-token ratio is computed for each concept per sentence, as well as for the entire text. As complex concepts tend to be more specific, the number of WordNet hypernyms and hyponyms is also taken into account. The WordNet number of senses for each concept is computed to account for polysemy and abstractness. Some concepts such as e.g. “love”, may be complex, yet they are mentioned so frequently that their complexity is not perceived anymore. To account for this, the frequency per one million words for each mention of a concept is taken from the British National Corpus [22]. Finally, for each concept, the number of ConceptNet relatives is counted if its weight, that is its confidence score, is greater than or equal to 0.5.

3.3 Human Annotation

For this task, a total of 247 sentences are taken from “The Great Gatsby” by F. Scott Fitzgerald, “Moby Dick” by Herman Melville, as well as “Persuasion” by Jane Austen. Those 247 sentences contain a total of 7978 tokens, 1687 of which are concepts. All of these are manually annotated by one human annotator according to a set of annotation guidelines. All concepts are judged in terms of their conceptual complexity where “0” denotes a conceptually simple concept, while “1” means that the concept is conceptually complex.

3.4 Model

Using h2o’s AutoML [23], classification is attempted on all 1687 concepts. We run three subsets of the data, the first containing all features, while the second and third subsets each exclude either those features derived from WordNet or those derived from ConceptNet. The leading model in all three cases is a stacked model comprised of a deep learning model stacked on top of a gradient boosting machine, a random forest and decision trees.

4 Results

The stacked model was built using all 1687 concepts of which 1373 were used for training while 157 were used respectively for validation and testing. At this point in time, the model built on the subset of the data without the WordNet features (hypernyms, hyponyms, number of senses, path similarity) performs best, achieving a mean squared error of 0.062 and a f1-score of 0.7083. When training the model with a subset not containing the ConceptNet features (number of relatives, relatedness), the results achieved were the worst among the three subsets.

Table 1. Results for stacked machine learning model.

Metric	All Features	w/o WordNet	w/o ConceptNet
MSE	0.0653	0.062	0.0648
RMSE	0.2555	0.249	0.2546
Accuracy	0.931	0.9141	0.8865
Precision	0.75	0.6071	0.5224
Recall	0.5	0.85	0.875
F1 Score	0.6	0.7083	0.6542

Out of the 157 concepts in the test set, 15 were judged to be conceptually complex by the human annotator. While both the model built on all features, as well as the model built on the subset without ConceptNet features found around the same amount of complex concepts, 16 and 14, they included many false positives. In contrast, the model without WordNet features classified a total of 22 concepts as complex, yet included more actually complex concepts.

5 Discussion

As this is a work-in-progress paper and, as of yet, there are no finalised results, the discussion will be primarily focused on the weaknesses of this methodology rather than relating the findings to previous works in the field.

So far, due to computational restraints, the dataset used was fairly small. In the future, we plan on increasing the size of the dataset. Furthermore, the increased dataset would need to be annotated by more than one annotator in order to decrease the possibility of the one annotator having a bias that is being transferred onto the data.

In terms of feature engineering, other knowledge graphs and ontologies should be explored to possibly replace those features based on the WordNet ontology. A big shortcoming of WordNet is that its depth varies in different parts. This leads to some concepts having less hyper -or hyponyms than others and, hence, the possibility of these being falsely classified as simple when they are in fact complex.

Furthermore, the predictions on the test set show that each of the three models has a slight tendency to falsely classify longer words as more complex. Future experiments should experiment with excluding the word-length feature to see whether this can improve results.

6 Conclusion

In this paper, we proposed a new definition for conceptual complexity. We extracted concepts from a text and use a variety of pairwise, as well as single features, to classify them as either simple or complex. It is proposed that our research forms part of text simplification systems in order to improve output.

While the estimation of conceptual complexity can be a good precursor for text simplification systems, it could also be useful for various other natural language processing tasks. Taking Machine Translation (MT) as an example, conceptual estimation could be performed as a step in the MT pipeline for under-resourced language.

For future work, it would be interesting to experiment with more features involving different types of knowledge graphs as well as including more pair-based features.

References

1. Adult Literacy, <https://www.literacytrust.org.uk/parents-and-families/adult-literacy/>. Last accessed 14 May 2021
2. Siddharthan, A.: A survey of research on text simplification. *ITL - International Journal of Applied Linguistics* **165**(2), 159–298 (2014)
3. Paetzold, G. H., Specia, L.: A survey on lexical simplification. *Journal of Artificial Intelligence Research* **60**, 549–593 (2017)
4. Saggion, H.: Automatic Text Simplification. *Synthesis Lectures on Human Language Technologies* **10**(1), 1–137 (2017)

5. Sikka, P., Singh, M., Pink, A., Mago, V.: A survey on text simplification. *Journal of the Association for Computing Machinery* **37**(4), (2020)
6. Dolch, E. W.: Fact burden and reading difficulty. *The Elementary English Review* **16**(4), 135–138 (1939)
7. Štajner, S., Hulpuş, I.: Automatic assessment of conceptual text complexity using knowledge graphs. In: *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 318–330. ACL, Santa Fe, New Mexico, USA (2018)
8. Štajner, S., Hulpuş, I.: When shallow is good enough: Automatic assessment of conceptual text complexity using shallow semantic features. In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1414–1422. ACL, Marseille, France (2020)
9. Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics* **165**(2), 97–135 (2014)
10. Daneman, M., Carpenter, P. A.: Individual differences in working memory and reading. *Journal of verbal learning and verbal behaviour* **19**(4), 450–466 (1980)
11. Shiffrin, R. M., Schneider, W.: Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review* **84**(2), 127–190 (1977)
12. Palladino, P., Cornoldi, C., De Beni, R., Pazzaglia, F.: Working memory and updating processes in reading comprehension. *Memory and cognition* **29**(2), 344–354 (2001)
13. Vajjala, S., Meurers, D.: Assessing the relative reading level of sentence pairs for text simplification. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 288–297. ACL, Gothenburg, Sweden (2014)
14. Vajjala, S., Meurers, D.: Readability-based sentence ranking for evaluating text simplification. In: *arXiv preprint, arXiv:1603.06009* (2015)
15. Ambati, B. R., Reddy, S., Steedman, M.: Assessing relative sentence complexity using an incremental CCG parser. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1051–1057. ACL, San Diego, California (2016)
16. Graesser, A. C., McNamara, D. S., Louwerse, M. M., Cai, Z.: Coh-Metrix: Analysis of text on cohesion and language. *Behaviour research methods, instruments, & computers* **36**(2), 193–202 (2004)
17. Narayan, S., Gardent, C.: Hybrid simplification using deep semantics and machine translation. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 435–445. ACL, Baltimore, Maryland (2014)
18. Štajner, S., Glavaš, G.: Leveraging event-based semantics for automated text simplification. *Expert systems with applications* **82**, 383–395 (2017)
19. Oram, P.: *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts (1999)
20. Speer, R., Chin, J., Havasi, C.: ConceptNet5.5: An Open Multilingual Graph of General Knowledge. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4444–4451. AAAI, San Francisco, California (2017)
21. Lewis, M. L., Frank, M. C.: The length of words reflects their conceptual complexity. *Cognition* **153**, 182–195 (2016)
22. The British National Corpus, version 3. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/> (2007)

23. LeDell, E., Poirier, S.: H2O AutoML: Scalable Automatic Machine Learning. In: 7th ICML Workshop on Automated Machine Learning. AutoML (2020)