Comparative judgments are more consistent than binary classification for labelling word complexity

Sian Gooding, Ekaterina Kochmar, Alan Blackwell
Dept of Computer Science and Technology
University of Cambridge
{shg36|ek358|afb21}@cam.ac.uk

Advait Sarkar
Microsoft Research
Cambridge
advait@microsoft.com

Abstract

Lexical simplification systems replace complex words with simple ones based on a model of which words are complex in context. We explore how users can help train complex word identification models through labelling more efficiently and reliably. We show that using an interface where annotators make comparative rather than binary judgments leads to more reliable and consistent labels, and explore whether comparative judgments may provide a faster way for collecting labels.

1 Introduction

In this paper, we address the use of machine learning (ML) for natural language readability assessment concerned with the identification of factors that affect a reader's understanding, reading speed and level of interest (Dale and Chall, 1949). We focus on *lexical simplification*, which aims to adapt text by replacing contextually complex words with more accessible meaning-equivalent alternatives: e.g. replacing *ameliorate* with *improve* in the context like *"They aimed to* **ameliorate** → **improve** *the situation."* Lexical simplification can be framed as a two step procedure, where the algorithm needs to first identify which words (or more specifically word senses) in context require simplification, and then replace them with simpler alternatives. The first step is commonly referred to as *complex word identification* (CWI) (Shardlow, 2013).

In supervised ML, algorithms are trained using data that is labelled according to a target concept (Kulesza et al., 2014). In the CWI task, the concept is word complexity in context, which for a human reader may combine multiple factors that a machine tries to learn from the data. Labelling of large data sets is time-consuming and costly, and often carried out using crowd-sourcing platforms such as Amazon Mechanical Turk (Paolacci et al., 2010). For the CWI task, crowd-source workers have in the past been employed to identify which words within a training dataset are complex: for example, given a sentence *"They aimed to ameliorate the situation"*, the annotators might identify *ameliorate* as complex. Labelled datasets collected this way are then used to train a model that can predict previously unseen words' complexity. Prior work on labelling of CWI datasets has found that annotation of word complexity is challenging, yielding relatively low levels of inter-annotator agreement such as $\alpha = 0.244$ (Paetzold and Specia, 2016) and $\kappa = 0.398$ (Specia et al., 2012).

In this paper, we show that representing the concept of word complexity in a continuous manner results in higher inter-annotator agreement than using binary labels. In particular, we investigate the following hypothesis:

Hypothesis 1 (H1): *Do comparative judgments for CWI lead to higher inter-annotator agreement and higher quality labelled data than binary judgments?*

Furthermore, this paper poses the following questions regarding the general setting of the CWI annotation experiments:

1. Does controlling for the homogeneity of the group of annotators with respect to their age, education level and native language contribute to higher agreement?

2. Can comparative judgments be made in a significantly shorter period of time than binary judgments for word complexity?

2 Background

2.1 Collecting Complex Word Labels

CWI is an essential first step in the lexical simplification pipeline, and has recently received signif-

Proceedings of the 13th Linguistic Annotation Workshop, pages 208–214
Florence, Italy, August 1, 2019.

icant attention (Shardlow, 2013). However, there are few labelled datasets suitable for CWI training, and those that exist have a number of drawbacks:

- The homogeneity of the annotator group is usually not controlled for, meaning that labels are provided by individuals with various backgrounds, conflating factors such as age, native language and education. We believe that it is important to clearly define and control for such factors, especially since the reading needs of different groups vary substantively;
- The annotation task is often presented as a binary decision, with annotators being asked to label each word as either complex or not. Intuitively, word complexity is expected to be a continuum, meaning that scalar or rank approaches should be more appropriate;
- Perhaps as a consequence of these two factors, the inter-annotator agreement for the labels is very low – lower than would be expected to support consistent empirical results when training ML algorithms (Cohen, 1968; Krippendorff, 2004; Bhowmick et al., 2008).

The first labelled CWI dataset was collected for the 2012 iteration of the SemEval Task 1 (Specia et al., 2012). This dataset was based on the data from McCarthy and Navigli (2007) which focused on word substitutions. The training set was annotated by 4 people while the test set was annotated by 5. In the labelling task, annotators were shown a short input text and a target word in English. For the target word, several possible substitutions were provided and annotators were asked to rank these substitutions according to their simplicity, e.g.:

(1) **Gold**: *clear, bright, light, well-lit*

Since the original words were provided as the input, this task was primarily focused on ranking substitution candidates rather than the CWI step. The inter-annotator agreement was measured using Cohen's Kappa coefficient by calculating κ for each pair of annotators, and then averaging over all pairs to derive the final score. The κ value was 0.386 for the training and κ=0.398 for the test set. Cohen's suggested interpretation is that values in the range of $0.21-0.40$ represent minimal agreement (Cohen, 1968). Specia et al. (2012) report that, while these scores are low, they correctly reflect the highly subjective nature of the annotation task.

A second CWI dataset was collected and annotated for the 2016 SemEval Task 11 (Paetzold and Specia, 2016). Rather than aiming for a measure of word complexity, this task was designed to evaluate systems that would identify if target words in context were complex or not. Labels were collected from 400 non-native annotators aged between 18 and 66, having 45 language backgrounds. Annotators were asked to select words within a sentence that they considered to be complex. The total dataset contained 9,200 sentences.

Inter-annotator agreement was calculated using Krippendorff's α agreement coefficient (Hayes and Krippendorff, 2007) for each set of 10 sentences, and each sentence was annotated by 20 volunteers. Krippendorff's α is more appropriate than the κ coefficient for multiple annotators as well as binary and ordinal labelling schemes (Antoine et al., 2014). When interpreting the α coefficient, Krippendorff suggests that $\alpha \geq 0.667$ is the lowest conceivable limit for tentative conclusions (Krippendorff, 2004). F-scores showed significant difference in annotations ($p<0.05$) between the age bands. Paetzold and Specia (2016) reported the quantitative differences in the annotation by the different age and language proficiency groups of annotators, however these differences were not further investigated or controlled for.

Finally, the CWI dataset in the CWI 2018 shared task (Yimam et al., 2018) was based on the dataset by Yimam et al. (2017) and contained data representing three different genres: Wikipedia, professionally-written and non-professionally written news. Annotations for this data were collected from 20 annotators using the MTurk platform. To counteract previous low inter-annotator agreement, the annotators were incentivized to maximize agreement. The inter-annotator agreement (IAA) was not reported, meaning that this dataset cannot be directly compared with the other two datasets. However, it is worth noting that nearly 30% of the words were annotated as complex by only a single annotator, while only 1.1% were annotated as complex by all 20 annotators.

Data	*IAA Statistic*	*Interpretation*
2012	$\kappa = 0.386, 0.398$	minimal agreement
2016	$\alpha = 0.244$	inconclusive
2018	1% unanimous	idiosyncratic

Table 1: Standard of inter-annotator agreement in previous CWI datasets

In summary, Table 1 shows low values of the

statistical measures for each of the three previous datasets. High inter-annotator agreement is a key requirement for the usability of an annotated corpus, whereas inconsistent or noisy annotation contributes to poor classifier performance (Bhowmick et al., 2008).

2.2 Approaches to Labelling

It is widely understood that machine learning systems are limited by the quality of the labelled training data. One approach to improving the performance of such systems is to treat the human labeller(s) as a source of noise (Frénay and Verleysen, 2014) who can be modelled statistically (Yan et al., 2010) in order to more accurately identify an underlying ground truth. Noise estimation can be improved if multiple labels are obtained for each item in the training set in order to model inconsistency (Ipeirotis et al., 2014), or if a distribution of label values can be used as a basis for rejecting outliers (Brodley and Friedl, 1999). However, these approaches presume that there is a single correct label for each data point. For our task of word complexity, different reports of complexity may be equally valid for different raters, which means that rather than a single underlying ground truth, the concept itself is individually variable.

Several of the human factors elements can be addressed through the use of pairwise comparison, where labellers make relative judgments to compare training items, rather that attempting to characterize each item independently against an abstract conceptual category, for which they are expected to have a stable definition and associated membership criteria. In the context of labelling, comparative judgments are used to compare how well the training items correspond to the required concept. Carterette et al. (2008) demonstrate that this method can facilitate judgments for information retrieval applications. Comparative judgments have also been used in gamified labelling (Bennett et al., 2009), where cooperating players reduce the set of alternative items until agreement is reached.

Recent work has looked into the application of comparative judgments to labelling as opposed to assignment of categorical values and scores on a scale (Simpson et al., 2019; Yang and Chen, 2011; Kingsley and Brown, 2010). Simpson et al. (2019) note that comparative judgments are suitable for abstract linguistic properties, whose nature can cause inconsistencies in the assigned numerical scores. In this work, we assume that word complexity is an instance of such abstract linguistic property. In addition, it has been showed before that comparative labelling allows a total sorting of items and can reduce the time taken to label a dataset (Yang and Chen, 2011; Kingsley and Brown, 2010; Kendall, 1948). In the context of CWI and text simplification systems, the relative nature of word complexity and comparative labels can be utilized to help the systems focus on the most complex words in text (Gooding and Kochmar, 2019).

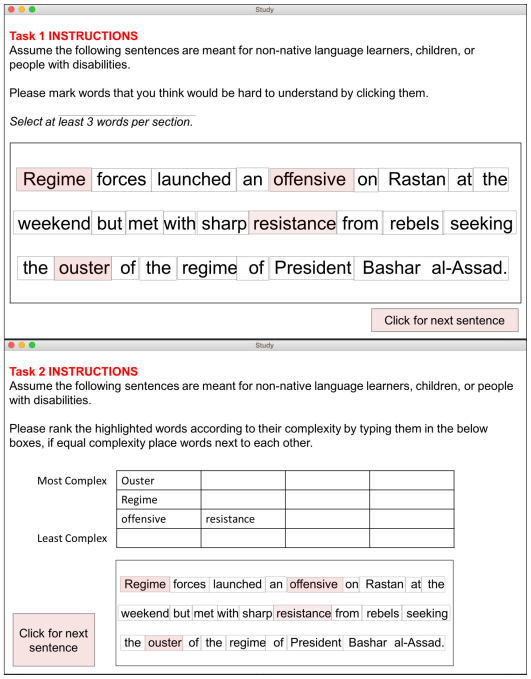

Figure 1: Labelling interfaces used in the study: Task 1 represents the binary annotation task, Task 2 – the ranking annotation task.

Finally, such factors as *interface design* (the simplicity of the interface and clarity of the instructions), *representation of target concept*, and *recruitment of annotators* (expertise or knowledge found in specific subgroups) are key to the reliability of annotation (Sarkar et al., 2016).

3 Study

In this paper we aim to study three points of interest: (1) whether controlling for such factors as age, level of education and native language of the annotator group in the task of complex word identification would yield higher inter-annotator agreement than reported in the previous studies; (2) whether modelling the labelling con-

cept as a comparative judgment better represents the concept of word complexity than categorical judgment, thus improving inter-annotator agreement measures; and (3) whether using comparative judgments is a more time-efficient way of labelling complex words. To investigate these, we performed a study using 30 annotators. The entire annotation process took approximately 25 minutes per participant. The participants were selected according to the following criteria: the same first language (English), the same level of educational background (graduate degree) and within a similar age range of 21-30. These initial criteria were motivated by the high availability of native speaking participants. In addition, by restricting the background of the participants, we aimed to show that homogeneity of the group of annotators can lead to higher inter-annotator agreement.

Two alternative interfaces, shown in Figure 1, were designed. We used a within-subjects design, in which each of the 30 annotators labelled 20 sentences (10 sentences per interface). The 20 sentences were extracted from the dataset of Yimam et al. (2017), which was chosen as the most reliable dataset for the task of CWI having yielded the best empirical results to date (Yimam et al., 2018). All sentences used in this study were selected from professionally written news, and were chosen to contain hard, medium and low complexity words as illustrated in Example 2. These words were selected using previous annotations reported for this dataset (Yimam et al., 2017). The proportion of annotators that mark a word as complex indicates the likelihood of the word being complex. We approximate the complexity strength using these measures, where the class boundaries are defined as: *hard* $\in [10, 20]$, *medium* $\in [6, 9]$, *low* $\in [1, 5]$.

(2) Hard: *politicizing* (14)
Medium: *warily* (9)
Low: *trip* (2)

This example shows words of different levels of complexity with the number of annotators that have marked them as complex (Yimam et al., 2017). Note that contrary to the study of Specia et al. (2012), where the annotators were asked to rank synonyms of approximately equal complexity, we ask them to rank words of different complexity. Having clear category differences has been shown to reduce cognitive load, thereby increasing labelling efficiency (Sarkar et al., 2016).

The first interface presented the labelling task as a classification exercise, allowing annotators to choose and label complex words by clicking on them. At least three words had to be selected before moving to the next sentence to ensure annotators' engagement in the task. The second interface presented the labelling task as a ranking exercise where words could be ordered according to their relative complexity. Words were ordered by re-entering them into a table with the position indicating the least to most complex words.

In both experiments, participants were asked to assume that the textual content was intended for a target audience of non-native language learners or people with reduced reading skills. To control for order effects, half the participants performed task 1 first, and half performed task 2 first.

4 Results

For the binary task, 62 distinct words from the 10 sentences were marked as complex by annotators. Two inter-annotator agreement measures are calculated for the binary and ranking tasks – Cohen's Kappa and Krippendorff's Alpha. The Kappa coefficient represents the average of scores across all pairs of raters for consistency with previous CWI studies. The inter-annotator agreement scores as well as the average labelling time per sentence are shown in Table 2.

	Comparative Judgment	*Binary Judgment*
Kappa Coefficient	0.6775	0.3937
Alpha Coefficient	0.6821	0.4960
Avg Time (s)	28.77	38.69

Table 2: Results of the study

Using the Kappa interpretations (Cohen, 1968), the comparative (ranking) labelling task has a *moderate* level of agreement, whereas the agreement in the binary annotation task is *minimal*, showing that the comparative judgment leads to a higher level of agreement than the binary categorisation judgment. At the same time, according to McHugh (2012), since the annotations obtained in our comparative judgment study result in a κ value above 0.60, they can be considered reliable. The α coefficient for the comparative judgment data also reflects this finding as it is above the required 0.667 threshold. This supports our hypothesis H1.

We note that the level of agreement in our binary annotation task is higher than the level of

agreement for the previously reported binary annotation tasks ($\alpha = 0.496$ vs $\alpha = 0.244$ in Paetzold and Specia (2016)). We also note that the level of agreement in our comparative judgment annotation task is higher than that in the previously reported studies ($\kappa = 0.6775$ vs $\kappa = 0.398$ in Specia et al. (2012)). We hypothesize that this is due to the more homogeneous group of annotators in our study, though this requires a more thorough investigation of the contributing factors and we leave the more controlled experimentation with various annotator backgrounds to the future.

We also note, that the average time per sentence for the ranking task is 9.92 seconds shorter than that for the binary task. Whilst this is partly expected due to the complex words being preselected, the annotator is still required to read and consider the words within context. These results suggest that ranking is a more efficient mechanism for collecting complex word annotations that results in a higher annotator reliability than traditional approaches. The statistical significance between annotation times was tested using an unpaired t-test and was found to be highly significant (p=0.001). We note that the current setting does not control for the differences in the two user interfaces or take into account the pre-annotation required to identify words in the ranking task, and leave more thorough experimentation on the comparative efficiency of the different approaches to labelling to the future.

5 Discussion and Future Work

This study demonstrates the advantage of annotating datasets using comparative judgments rather than binary classifications, both for efficiency and accuracy. Comparative labels are used relatively rarely in ML research at present, but our results suggest that this may be a more reliable basis for training such models in future, especially where the phenomenon to be modelled relies on human experience (Simpson et al., 2019).

A further advantage of constructing rankings rather than classifications is that we are able to infer additional labels without the need for further annotation, by using a pre-labelled framework (Sarkar et al., 2016). In particular, whereas we only get binary labels for the words in a binary setting, the relative ranking can be extended to the full dataset, thus increasing the size of the labelled data without additional effort. A number of methods for learning total sorting from sparsely annotated data have been proposed in the literature (Simpson and Gurevych, 2018; Marley and Louviere, 2005; Thurstone, 1927).

Our results also show higher agreement coefficients for both binary and relative judgment tasks when compared to previously collected datasets. This supports the case that the concept of word complexity, and thus the level of agreement, is aligned between individuals that share a common background, as for our sample. This emphasizes the importance of considering the annotator group carefully when constructing annotated training corpora, or carrying out labelling experiments. This paper sets the benchmarks for the CWI annotation experiments with a homogeneous group of native speaking annotators using interfaces for collecting comparative and binary judgments. The future steps for this research include: (1) more thorough investigation of effects of annotator group homogeneity on the inter-annotator agreement, and (2) more detailed study of the efficiency of the comparative judgments as opposed to binary judgments.

Finally, although in this work we have focused on the CWI task, our results are potentially applicable to other natural language tasks where specific user experiences like simplicity must be modelled as an ordering so that they can be optimized or personalized.

Acknowledgments

The second author is grateful to Cambridge English for supporting this research via the ALTA Institute.

References

Jean-Yves Antoine, Jeanne Villaneau, and Anas Lefeuvre-Halftermeyer. 2014. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation. *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*.

Paul N. Bennett, David Maxwell Chickering, and Anton Mityagin. 2009. Learning consensus opinion: mining data from a labeling game. In *Proceedings of the 18th international conference on World wide web*, pages 121–130. ACM.

Plaban Kr Bhowmick, Pabitra Mitra, and Anupam Basu. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 58–65. Association for Computational Linguistics.

Carla E. Brodley and Mark A. Friedl. 1999. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, pages 131–167.

Ben Carterette, Paul N. Bennett, David Maxwell Chickering, and Susan T. Dumais. 2008. Here or there preference judgments for relevance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4956 LNCS:16–27.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Edgar Dale and Jeanne S. Chall. 1949. The concept of readability. *Elementary English*, 26(1):19–26.

Benoit Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Sian Gooding and Ekaterina Kochmar. 2019. Complex Word Identification as a Sequence Labelling Task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Panagiotis G. Ipeirotis, Foster Provost, Victor S. Sheng, and Jing Wang. 2014. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery*, 28(2):402–441.

Maurice G. Kendall. 1948. *Rank Correlation Methods*. Griffin, Oxford, UK.

David C. Kingsley and Thomas C. Brown. 2010. Preference Uncertainty, Preference Learning, and Paired Comparison Experiments. *Land Economic*, 86:530–544.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured labeling for facilitating concept evolution in machine learning. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 3075–3084.

Anthony A. J. Marley and J. Louviere, Jordan. 2005. Some probabilistic models of best, worst, and best-worst choices. *Journal of Mathematical Psychology*, 49:464–480.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, 22(3):276–282.

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

Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419.

Advait Sarkar, Cecily Morrison, Jonas F. Dorn, Rishi Bedi, Saskia Steinheimer, Jacques Boisvert, Jessica Burggraaff, Marcus D'Souza, Peter Kontschieder, Samuel Rota Bulò, Lorcan Walsh, Christian P. Kamm, Yordan Zaykov, Abigail Sellen, and Siân Lindley. 2016. Setwise comparison: Consistent, scalable, continuum labels for computer vision. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 261–271. ACM.

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.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting Humorousness and Metaphor Novelty with Gaussian Process Preference Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*. Association for Computational Linguistics.

Edwin Simpson and Iryna Gurevych. 2018. Finding Convincing Arguments Using Scalable Bayesian Preference Learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 Task 1: English Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, SemEval '12, pages 347–355, Stroudsburg, PA, USA. Association for Computational Linguistics.

Louis L. Thurstone. 1927. A law of comparative judgement. *Psychological Review*, 34:273–286.

Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer G. Dy. 2010. Modeling annotator expertise: Learning when everybody knows a bit of something. In *International conference on artificial intelligence and statistics*, pages 932 – 939.

Yi-Hsuan Yang and Homer H. Chen. 2011. Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:762–774.

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)*, volume 2, pages 401–407.