CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting

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

This paper presents the winning systems we submitted to the Complex Word Identification Shared Task 2018. We describe our best performing systems' implementations and discuss our key findings from this research. Our best-performing systems achieve an F_1 score of 0.8736 on the NEWS, 0.8400 on the WIKINEWS and 0.8115 on the WIKIPEDIA test sets in the monolingual English binary classification track, and a mean absolute error of 0.0558 on the NEWS, 0.0674 on the WIKINEWS and 0.0739 on the WIKIPEDIA test sets in the probabilistic track.

1 Introduction

Poor reading comprehension often caused by the presence of complex technical terms can have serious practical consequences (Dubay, 2004). Although proper text simplification requires a wide range of transformations, it has been shown that application of lexical simplification (LS) techniques alone improves reader understanding and information retention (Leroy et al., 2013). Complex Word Identification (CWI) is concerned with automated identification of words that might present challenge for the target readers and should thus be simplified (Shardlow, 2013a). Early studies on LS (Carroll et al., 1999; Devlin and Tait, 1998) do not consider CWI as part of the simplification pipeline, but recent studies argue that simplification systems benefit from applying CWI as the first step in the LS pipeline (Shardlow, 2014; Paetzold and Specia, 2016b). Inadequate identification of complex words in text might result in an overly difficult text if many potential candidates are missed, or in meaning distortion if many simple words are falsely identified as complex.

CWI can not only be used as a component of LS systems, but also as a stand-alone application within intelligent tutoring systems for second language learners or in reading devices for people with low literacy skills. For instance, Nation (2006) shows that at least 95% of text should be familiar to the reader in order for them to understand the content. A CWI system can help identify the unfamiliar words and provide readers with their definitions even when simpler alternatives are not available. This has the potential to help a wide variety of target reader groups, including general readers of technical texts (Feng, 2008).

Following the SemEval 2016 shared task (Paetzold and Specia, 2016c), the Shared Task 2018 frames CWI as the process of identifying words that are difficult for a given target population (for example, non-native speakers of English) based on the annotation from a sample of that target population (Yimam et al., 2018). We overview the related work in the field in Section 2 and discuss the CWI shared task framework in Section 3. We have participated in the binary and probabilistic classification tasks in the monolingual English track, and scored first in the binary setting on all three data sources, as well as on two out of three data sources in the probabilistic setting. Section 4 presents the implementation details of our systems including features and methods used. In Section 5 we present the results obtained with our systems, and discuss the key findings. Finally, we outline future directions for this research in Section 6.

2 Related Work

The earliest studies that address CWI as an independent task are related to the medical domain: Zeng et al. (2005) predict medical term familiarity based on term occurrence, and show that individualised assessment is possible if the models consider readers' demographics. Elhadad (2006), in addition to corpus frequency, consider using familiarity features from the MRC Psycholinguistic

Database (Wilson, 1988) and the number of senses from WordNet (Fellbaum, 2005). Zeng-Treitler et al. (2008) improve on the previous methods using contextual information.

Some previous approaches to LS consider all words as potentially complex and try to simplify every word (Devlin and Tait, 1998; Thomas and Anderson, 2012; Bott et al., 2012). This has a number of undesirable effects, including radical changes in the original meaning and the dependence of the simplification process on the availability of alternatives. For instance, the results of Horn et al. (2014) show that such an approach is unable to find a simpler alternative for one third of the complex words in their dataset. Another type of approach introduces a threshold that is typically based on the word frequency (Zeng et al., 2005; Elhadad, 2006; Biran et al., 2011). Until recently (Shardlow, 2013b), the lack of shared data to compare different approaches to CWI has been one of the bottlenecks for this task.

The CW corpus of Shardlow (2013b) is based on the edit histories in Simple Wikipedia, and includes only the sentences where a single word is simplified. Paetzold and Specia (2016c) find that as much as 51.9% of the words in this corpus are annotated as complex by at least one of their annotators and conclude that non-native speakers of English might still find the simplified version of Wikipedia challenging. The quality of Simple Wikipedia and its usefulness for simplification research has been challenged before in Xu et al. (2015). Further experiments in Shardlow (2013a) show that a more resource-intensive threshold-based approach does not perform significantly differently on this dataset to a more naïve technique of simplifying everything, while an SVM classifier performs better in terms of precision but does so at the cost of a much lower recall. These findings inspired further research into classification-based approaches to CWI (Paetzold and Specia, 2016c).

The SemEval 2016 shared task on CWI combines the data from the CW corpus of Shardlow (2013b), the LexMTurk corpus of Horn et al. (2014) and the Simple Wikipedia corpus of Kauchak (2013), all of which rely on Simple Wikipedia data. A set of 400 non-native speakers annotated the content words in the data as simple or complex. The information about annotator's age, native language (L1), education and level of language proficiency has been collected, but has not been used in the task. The final dataset has a bias towards annotation provided by the non-native speakers of upper levels of language proficiency and, potentially as a result of that, only about 11% of word types (and 3% of word tokens) are annotated as complex (Paetzold and Specia, 2016c). The results of the shared task show that simpler features based on word frequency (Konkol, 2016; Wróbel, 2016; Zampieri et al., 2016) and word presence in certain lexicons (Mukherjee et al., 2016; Wróbel, 2016), work best. A number of systems performing best in terms of G-score used various ensemble-based approaches (Paetzold and Specia, 2016d; Ronzano et al., 2016; Mukherjee et al., 2016; Zampieri et al., 2016). The systems that performed best in terms of F-score used threshold-based approaches (Wróbel, 2016; Malmasi et al., 2016) and frequency features (Malmasi and Zampieri, 2016).

In their analysis of the SemEval 2016 shared task, Zampieri et al. (2017), similarly to Paetzold and Specia (2016c), show that an ensemble of all systems does not outperform the best system or an ensemble of a few best-performing systems. The use of an oracle of the 3 best-performing systems sets the upper bound at 0.60 F-score for the identification of complex words and at 0.98 F-score for the identification of simple words. They also show that the systems more reliably identify those complex words that are annotated as such by the majority of human annotators, arguing that lexical complexity should be seen as a continuum on a spectrum rather than a binary value.

3 CWI Shared Task 2018 Setup

The CWI shared task uses the data from Yimam et al. (2017b), and approaches CWI from two perspectives: under the *binary* (`bin`) view, a word can be either complex or simple, and in the *probabilistic* (`prob`) setting a word receives a score in the range of $[0.0, 1.0]$ reflecting the proportion of annotators that consider the word complex. In this section, we briefly overview the CWI shared task 2018 framework, discuss the data and the annotation, and analyse the challenges the CWI systems are presented with in this task.

3.1 Data

Unlike the previous datasets that rely on the use of Wikipedia and Simple Wikipedia, the CWIG3G2 dataset of Yimam et al. (2017a) uses texts of 3

different genres: professionally written news articles (NEWS), amateurishly written news articles (WIKINEWS), and WIKIPEDIA articles. The dataset includes annotation for content words as well as for phrases. The annotation for the English data is collected from both native and non-native speakers of English. Table 1 presents the statistics on the number of words (*w*) and phrases (*ph*) in the training (*train*), development (*dev*) and *test* subsets of the News (NEWS), WikiNews (WINS) and Wikipedia (WIKI) datasets.

Data	Train	Dev	Test
NEWS (w)	11,949	1,502	1,813
NEWS (ph)	2,053	262	282
WINS (w)	6,780	776	1,138
WINS (ph)	966	94	149
WIKI (w)	4,833	606	750
WIKI (ph)	718	88	120
TOTAL	27,299	3,328	4,252

Table 1: Number of instances

3.2 Annotation

The annotation was performed using the Amazon Mechanical Turk platform. The set of annotators comprised 10 native and 10 non-native speakers of English. They were presented with text paragraphs and were asked to select up to 10 lexical items that they found complex. The lexical items included content words (e.g., nouns, verbs, adjectives and adverbs) and phrases up to 50 characters in length. Additional information about the annotators, such as their language proficiency, was collected but was not used in the task.

By allowing the annotators to select phrases as well as individual words, Yimam et al. (2017a) created a more practically useful dataset. By presenting the annotators with whole paragraphs, they replicated a realistic scenario in which words are interpreted in context. By not preselecting target lexical items, they avoided introducing the bias into the annotation, although it may be argued that the limit of 10 lexical items per paragraph restricted the selection options. Finally, since the annotations are provided by both native and non-native speakers, this allows Yimam et al. (2017a) to explore to what extent the needs of non-native speakers can be estimated based on the needs of a wider target population. The analysis in Yimam et al. (2017a) shows that there are quantitative differences between the annotation provided by the native and non-native speakers, and between the three genres. Further experiments show that the system trained on native speakers' annotations performs better than the system trained on non-native speakers' annotations, both on native and non-native data. Yimam et al. (2017a) also note that the inter-annotator agreement between native speakers is higher than between non-native speakers, which might be due to the fact that, unlike non-native annotators, native speakers share L1 and are of relatively similar language proficiency level. At the same time, these results suggest that the annotation provided by the native speakers can be used to predict the simplification needs of the non-native speakers as well.

The shared task relies on two types of annotation: under the `bin` setting that words and phrases are annotated as complex (label `1`) if at least one of the 20 annotators annotated them as such, and simple (label `0`) otherwise; and under the `prob` setting that words and phrases receive a label in the range between $[0.0, ..., 1.0]$, with a step of 0.05, reflecting the proportion of annotators who found the lexical item complex.

Table 2 presents the distribution of simple and complex words in the dataset. We present the label break-down in terms of label percentages across the genres (NEWS, WINS, WIKI) and subsets of data (*tr* for training, *dev* for development, and *ts* for test sets). Due to space limitations, for the `prob` setting we present only the percentage of cases annotated as simple ($0_{bin} = 0.0_{prob}$), annotated as complex by a single annotator (0.05_{prob}) and by all 20 annotators (1.0_{prob}).

Data	0_{bin}	1_{bin}	0.05_{prob}	1.0_{prob}
NEWS_{tr}	60.41	39.59	13.52	0.39
NEWS_{dev}	60.54	39.46	13.83	0.28
NEWS_{ts}	61.72	38.28	12.70	0.29
WINS_{tr}	58.48	41.52	16.25	0.17
WINS_{dev}	59.43	40.57	14.25	0.11
WINS_{ts}	57.58	42.42	16.71	0.16
WIKI_{tr}	55.07	44.93	16.66	0.52
WIKI_{dev}	51.15	48.85	19.31	0.14
WIKI_{ts}	49.54	50.46	18.62	0.23

Table 2: Annotation labels break-down (%)

These figures demonstrate that: (1) there is a quantitative difference in the annotation across the three genres, with NEWS being the easier to understand for the annotators (38.28% to 39.59% com-

plex words) and WIKI being the most complex (44.93% to 50.46% complex words), which suggests that systems might perform better if trained and tested within the same genre; (2) the distribution of complex and simple words across training, development and test subsets is consistent for NEWS and WINS with a difference in label distribution of no more than 2.5% – this suggests that the systems for these two genres might generalise better than the one for WIKIPEDIA; (3) about 1/3 of the complex word annotation comes from a single annotator finding a word complex, while the cases where all 20 annotators agree that the word is complex comprise less than 1% in all subsets. Furthermore, we have identified the following challenges presented by the dataset:

Context-specific annotation: Since the lexical items were presented to the annotators in a variety of contexts, the item might have received different annotation depending on the context. Between 3% and 10% of lexical items in the binary setting received different annotation, and in the probabilistic setting a number of words received a wide range of labels: e.g. the labels for *observatory* range from 0.0 to 0.95, and for *tragedy* from 0.0 to 1.0. There are several possible reasons for this effect:

- surrounding context might help or impede understanding of a target word;
- the word might be used in a rare sense;
- the data might show a sequential bias effect (Mathur et al., 2017).

Consider the following example from the WIKI training set:

(1) Beethoven's *Symphony*$_{0.6}$ No.7, Bruckner's *Symphony*$_{0.1}$ No.6 and Mendelssohn's *Symphony*$_{0.0}$ No.4 comprise a nearly complete list of *symphonies*$_{0.3}$ in this key in the Romantic era.

The first occurrence of the word *symphony* is annotated as complex by 12 annotators, the second one by 2 while the third one is not considered complex by any annotators. This might suggest that (1) by the third occurrence of the same word the annotators perceive the word as familiar, (2) some of them found it unnecessary to annotate a word multiple times, (3) given the restriction of 10 complex items per paragraph they prioritised other words. The annotation of *symphonies* illustrates that the annotators might find different morphological forms of the same word challenging.

Phrase annotation: The annotators were allowed to select phrases of up to 50 characters in length. At the same time, the component words within the phrase might have been annotated as complex independently by other annotators. This results in cases like (2), where the phrase complexity is a derivative of the component word complexities, as well as (3) where the phrase annotation is independent of the component words:

(2) *future*$_{0.05}$ ∪ *generations*$_{0.25}$ = *future generations*$_{0.15}$

(3) *traditional*$_{0.2}$ ∪ *connection*$_{0.0}$ ∪ *country*$_{0.05}$ ≠ *traditional connection to that country*$_{0.0}$

Annotation of proper nouns : Proper nouns received a variety of labels: e.g., from 0.0 to 0.45 for *Eurozone*, 0.0 to 0.05 for *Barack*, 0.05 to 0.3 for *Brexit*, and from 0.0 to 0.05 or 0.1 for a number of geographical locations like *Copenhagen*, *Estonia*, *Hungary*, *Warsaw*, etc. The annotation in such cases depends more on world knowledge than on the properties of the words *per se*.

3.3 Evaluation

The systems in the `bin` setting are evaluated using F-score. The systems in the `prob` setting are evaluated using *mean absolute error* (*MAE*) which estimates the average difference between the values in the gold standard and values predicted by the system across all test instances.

4 CAMB systems

This section describes the implementation details of the `CAMB` machine learning framework applied to the shared task data sets.

4.1 Features

The set of features employed in our experiments are based on the insights from the CWI shared task 2016 (Paetzold and Specia, 2016d). In addition, we incorporate (1) the number of words grammatically related to the target one, (2) a range of psycholinguistic features from the MRC Psycholinguistic Database (Wilson, 1988), (3) CEFR levels (Council of Europe, 2011) extracted from the Cambridge Advanced Learner Dictionary

(CALD),[1] and (4) the use of Google N-gram word frequencies sourced using the *Datamuse* API.[2]

4.1.1 Word N-gram and POS

The target word and its syntactic class are represented as matrices of token counts. For words, the token counts represent the whole vocabulary as well as character-based bi-grams contained within the words (*N-gram*). The part of speech tags (*POS*) each correspond to a value within the matrix. The syntactic class of the word is obtained by performing part of speech tagging on all sentences containing target words using the NLPCore pipeline (Manning et al., 2014).

4.1.2 Lexical Features

These features are based on the lexical information about the target word and include:

- *Word length (Len)*: the number of characters in the word.
- *Number of syllables (Syll)*: the syllable count for the target word, collected using the *Datamuse* API.
- *WordNet Features*: number of senses (*Syn*), number of hypernyms (*Hyper*) and hyponyms (*Hypo*) for the word's lemma from WordNet (Fellbaum, 2005).

4.1.3 Dependency Parse Relations

The data was parsed using the NLPCore pipeline, and the number of dependency relations for the target word are extracted and used as a feature (*DepNum*).

4.1.4 Lexicon-Based Features

All but the the last in the following list of features are binary features indicating the presence of the word within a lexicon. CALD returns a CEFR level of the target word on the scale $[1, 6]$:

- *SubIMDB*: a list produced using the SubIMDB corpus (Paetzold and Specia, 2016a). The word frequency in the subtitles from the '*Movies and Series for Children*' section is calculated, and the top $1,000$ words are included in this list.
- *Simple Wikipedia (SimpWiki)*: a list of the top $6,368$ words contained in the Simple Wikipedia (Coster and Kauchak, 2011).
- *Ogden's Basic English*: a list of $1,000$ words from Ogden's Basic English list (Ogden, 1968).
- *Cambridge Advanced Learners Dictionary (CALD)*: the entries contained in the Cambridge Advanced Learner's Dictionary with their CEFR levels.

4.1.5 Word Frequency

The frequency of the target word (*Freq*) is estimated using the Google dataset of syntactic n-grams (Goldberg and Orwant, 2013).

4.1.6 MCR Features

We extract the psycholinguistic features of the target words from the MCR Psycholinguistic Database (Wilson, 1988). As the coverage of this database is relatively low, if a target word is not in the dataset we use a $null$ value.

- *Word familiarity rating (FAM)* in the range of $[100, 700]$ is based on a combination of 3 sets of familiarity norms: Pavio (unpublished), Toglia and Battig (1978) and Gilhooly and Logie (1980).
- *Number of phonemes (NPHN)*
- *Thorndike-Lorge written frequency (TLFRQ)* – the frequency of occurrence derived from Thorndike and Lorge (1944).
- *Imageability rating (IMG)*, representing the ease of associating the word with an image, is derived from the same combination of sets as the familiarity rating.
- *Concreteness rating (CNC)* represents the degree to which the concept denoted by a word refers to a perceivable entity based on the norms of Gilhooly and Logie (1980).
- The *number of categories (*KFCAT*)*, *samples (*KFSMP*)* and *written frequency (*KFFRQ*)* are derived from Kučera and Francis (1967).
- *Age of acquisition (*AOA*)* is based on the norms of Gilhooly and Logie (1980), multiplied by 100 to produce a number in the range of $[100, 700]$ (min 125, max 697, mean 405, SD 120).

4.2 Method

Below we outline how the features are incorporated into the machine learning frameworks for the classification and regression tasks. We use distinct approaches to model word and phrase complexity.

[1] Publicly available through http://www.englishprofile.org/wordlists

[2] https://www.datamuse.com/api/

4.2.1 Binary Classification Approach

As a wide range of heterogeneous features are employed by both the classification and regression systems, a feature union pipeline is applied. We use the `sklearn` machine learning framework.[3] The numerical features are normalized using a Standard Scaler, which subtracts the mean and scales the data to unit variance. Text-based features are represented as a matrix of token counts using a Count Vectorizer component.

Experiments on the development set confirm the findings of Paetzold and Specia (2016c) that the best performing classification algorithms for this task are ensemble-based techniques. Of these, the boosting classifier `AdaBoost` with 5000 estimators achieves the highest results, followed by the bootstrap aggregation classifier `Random Forest`. For the WIKIPEDIA and NEWS datasets, the best performance is attained using AdaBoost. However for the WIKINEWS an ensemble voting classifier that combines both the AdaBoost and Random Forest classifiers with equal weightings gives the highest F-Score.

4.2.2 Experimental Setup

Feature Selection

The effectiveness of features varies according to the data set classified. For the WIKINEWS and NEWS all aforementioned features are integrated into the systems. The feature set for WIKIPEDIA does not include MCR psycholinguistic features.

Training Data

The performance of the classifier also varies according to the genre of data used for training. The WIKIPEDIA and WIKINEWS are best classified when all available training data are used (i.e., NEWS, WIKINEWS and WIKIPEDIA combined), whereas the best results are achieved on the NEWS when the system is trained using the NEWS dataset only.

4.2.3 Probabilistic Classification Approach

The probabilistic setting uses the same set of features as the binary classification algorithms. We use the `Linear Regression` algorithm, and the lowest MAE values are achieved with the following settings: we use all features and all training data for the NEWS, all but MRC psycholinguistic features and all training data for the WIKINEWS, and a combination of the WIKIPEDIA and WIKINEWS training data and all but MRC psycholinguistic features for the WIKIPEDIA.

Since the gold standard labels for the probabilistic classification tasks lie in the range of $[0.0, 1.0]$ with a step of 0.05 reflecting the proportion of annotators, we round the classifier's prediction to the nearest value on this scale.

4.2.4 Phrase Complexity Prediction

Table 1 shows that there are a non-negligible amount of phrases in the data. We implement three binary classification approaches and one probabilistic classification approach to predict phrase complexity.

Binary Classification Techniques

- *CW presence*: Each word within the phrase is first classified using our word-based CW classifier. If the total number of complex words is above a pre-defined threshold then the phrase is marked as complex.

- *N-gram classifier*: The frequency of n-grams contained within phrases is obtained from the Corpus of Contemporary American English (Davies, 2009). An AdaBoost classifier is first trained using these frequencies as features, and then applied to classify new phrase instances.

- *Greedy approach*: The greedy baseline approach simply labels all phrases as complex.

Probabilistic Classification Techniques

For the probabilistic setting, we first apply our word-based CW regression classifier, and then derive the phrase complexity label as the mean of the complexity values within the phrase. Note that this technique helps us correctly predict the phrase complexity for cases similar to example (2) from Section 3, but not for cases similar to example (3).

5 Results

In this Section, we present and discuss the results obtained with the `CAMB` systems. The systems submitted to the shared task scored first in the binary classification English track on all three text genres, first on the WIKINEWS and WIKIPEDIA test sets and third on the NEWS test set in the probabilistic classification English track. Table 3 presents the results, with those that scored first in the shared task marked in bold.

[3] http://scikit-learn.org/stable/

5.1 Test Set Results

	Binary (F-Score)	Probabilistic (MAE)
NEWS	**0.8736**	0.0558
WINS	**0.8400**	**0.0674**
WIKI	**0.8115**	**0.0739**

Table 3: Test set results

The final test files across all genres contain a total of $3,701$ words and 551 phrases. Words are classified using the tailored approaches according to the dataset genre. In the shared task submission, phrases are independently classified using the greedy approach (see Section 4).

5.2 Analysis

Per-Genre Performance

Classification performance as well as feature relevance varies across the datasets. In the binary setting, the highest performance is obtained on the NEWS data when the system is trained on the NEWS data only. In the probabilistic setting, the system performs best on the NEWS data as well. Table 2 suggests that NEWS contains the lowest number of complex words, and Table 4 shows the total number of words, the number of unique words and the percentage of unique words within each genre.

	NEWS	WINS	WIKI
Total	$13,461$	$7,559$	$5,439$
Unique	$3,376$	$3,334$	$3,157$
%	25.08	44.10	58.44

Table 4: Unique words distribution

Table 4 suggests that the NEWS dataset contains the lowest number of unique words, which might be the effect of more restricted vocabulary used in professional news. As a result, the classifier is likely to have multiple exposure to the same word (albeit in different contexts) during training. At the same time, WIKIPEDIA with its 58.44% has the highest ratio of unique words, which might be due to the fact that it covers a very broad range of subjects. Note, that WIKIPEDIA is both more challenging for human annotators (highest percentage of complex words in Table 2) and the classifiers (lowest results in both settings in Table 3). This might explain why the classifiers benefit from training on multiple data sources in this case.

Our CWI systems are context-independent, which means that a word or a phrase receives the same complexity label regardless of a particular context of use. E.g., all three occurrences of the word *Symphony* in the example (1) from Section 3 would receive the same complexity label from our system. This limitation is the biggest source of error for the NEWS dataset (88.94% of the misclassified words in the NEWS test set have multiple labels in the data) and the WIKINEWS dataset (61.31%), while the proportion of such cases in the WIKIPEDIA data is lower (52.78%) which might also be due to the higher ratio of unique words in the WIKIPEDIA data.

Phrase Classification Results

The CAMB submission to the shared task applies a simple *greedy* approach to the phrase classification. We run experiments with more informed approaches overviewed in Section 4 and evaluate whether these approaches improve performance. Table 5 presents the results obtained with the different approaches to the phrase classification in the binary setting. The results of the system submitted to the shared task are marked in bold.

Data	Acc	P	R	F-Score
CW pres.	0.6987	0.8049	0.8231	0.8139
N-gram	0.8004	0.8015	0.9977	0.8889
Greedy	**0.8004**	**0.8004**	**1.000**	**0.8891**

Table 5: Binary classification results for the phrase classification in the test set

The results suggest that the more linguistically informed n-gram classifier is capable of achieving results similar to the baseline greedy approach that simply labels all phrases as complex. To test how it would score in the shared task, we re-run the experiments using the n-gram based phrase classifier, and report the results in Table 6.

Data	Acc	P	R	F-Score
NEWS	0.8535	0.7778	0.8641	0.8479
WINS	0.8423	0.8046	0.8297	0.8392
WIKI	0.8081	0.8254	0.7859	0.8080

Table 6: Test set results using n-gram phrase classifier

We note a drop of 3.13% for the F-score on the NEWS dataset, although the difference in the F-score on the other two datasets is less than 1%.

Features	NEWS	WINS	WIKI
N-grams	0.792	0.789	0.754
POS	0.033	0.035	0.046
Freq	0.029	0.029	0.043
Syn	0.020	0.027	0.013
FAM	0.016	0.008	0.019
Syll	0.013	0.021	0.018
KFSMP	0.012	0.010	0.008
SimpWiki	0.010	0.011	0.005
TLFRQ	0.010	0.009	0.011
CNC	0.009	0.010	0.009

Table 7: Gini coefficient for feature contribution

Individual Feature Contribution

We analyse the contribution of individual features to the classification framework. Table 7 reports the Gini coefficient for the top 10 informative features across different datasets. The Gini coefficient is defined as the total decrease in node impurity, weighted by the probability of reaching that node, averaged over all trees of the ensemble (Breiman, 2015).

We also note that the combination of all features achieves best results on the NEWS and WIKINEWS data sets, but the results on the WIKIPEDIA data decrease when the MCR Psycholinguistic Database features are included. We have noted above that one of the reasons for lower performance on the WIKIPEDIA data is due to the more diverse vocabulary. In addition to that, we note that the MCR Psycholinguistic Database contains values for 150,837 words, but some measures provide much lower coverage (see Table 8).

Measure	Coverage (words)
AOA	3,503
CNC	8,228
IMG	9,240
FAM	9,392
TLFRQ	25,308
KFCAT/SMP/FRQ	29,778
NPHN	38,438

Table 8: Number of feature instances covered by the MCR Database

As the WIKIPEDIA dataset has the largest proportion of unique words, it is likely that these features do not improve the classification accuracy due to their sparsity.

Performance Across Parts of Speech

Table 9 reports the results achieved by the binary classification algorithm on the different parts of speech in the test files. We include only content words in our analysis.

Data	Size	Acc	P	R	F
Total Test	3,701	0.86	0.82	0.79	0.85
Nouns	2,427	0.86	0.80	0.76	0.84
Verbs	718	0.84	0.83	0.81	0.84
Adjectives	435	0.88	0.86	0.86	0.87
Adverbs	111	0.91	0.89	0.92	0.91

Table 9: POS Classification Metrics

We note that nouns represent the largest proportion of all test items, while showing the lowest precision and recall. We hypothesise that one of the reasons for that might be the dependence of the noun annotation on the context and the context-independent nature of our classifiers. In addition, as we note in Section 3, the complexity of proper nouns largely depends on the world knowledge and is harder to model with a machine learning approach: 12.56% of misclassified instances in the NEWS data, 22.02% in the WIKINEWS and 22.92% in the WIKPEDIA are proper nouns.

6 Conclusion

In this paper, we have presented the implementation of the CAMB systems submitted to the CWI Shared Task 2018, and discussed the key challenges for the systems. Our systems scored first on three text genres in the binary classification track, and on two out of three genres in the probabilistic track. Further analysis of the performance identifies future directions for this research.

First of all, our systems are implemented in a context-independent way, while the context of use clearly affects the perception of word complexity. Future research will look into the ways to include contextual features into the machine learning framework. In addition, future work should investigate how phrase complexity is derived from individual word complexity scores.

Secondly, we believe that the notion of word complexity is dependent on a number of demographic factors such as one's level of education, L1 and level of language competence. These factors should be included both at the data annotation step and at the CW detection step.

Acknowledgments

We are grateful to Cambridge English for supporting this research via the ALTA Institute.

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:shortpapers*, pages 496–501, Portland, Oregon. Association for Computational Linguistics.

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, Mumbai. COLING.

Leo Breiman. 2015. Random forests leo breiman and adele cutler. *Random Forests-Classification Description*.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying Text for Language-Impaired Readers. In *Proceedings of the 9th Conference of the European Chapter of the ACL (EACL'99)*, pages 269–270, Bergen, Norway.

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 665–669. Association for Computational Linguistics.

Mark Davies. 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.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

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

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

The Council of Europe. 2011. Common European Framework of Reference for Languages: Learning, Teaching, Assessment.

Christiane Fellbaum. 2005. *Encyclopedia of Language and Linguistics, Second Edition*, chapter WordNet and wordnets. Oxford: Elsevier.

Lijun Feng. 2008. Text simplification: A survey. Technical report, CUNY.

Ken J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1, pages 241–247.

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

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st ACL*, pages 1537–1546.

Michal Konkol. 2016. Uwb at semeval-2016 task 11: Exploring features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1038–1041, San Diego, California. Association for Computational Linguistics.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Gondy Leroy, James E. Endicott, David Kauchak, Obay Mouradi, and Melissa Just. 2013. User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention. *Journal of Medical Internet Research (JMIR)*, 7(15).

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000, San Diego, California. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2016. Maza at semeval-2016 task 11: Detecting lexical complexity using a decision stump meta-classifier. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 991–995, San Diego, California. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2017. Sequence Effects in Crowdsourced Annotations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2860–2865.

Niloy Mukherjee, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2016. Ju_nlp at semeval-2016 task 11: Identifying complex words in a sentence. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 986–990, San Diego, California. Association for Computational Linguistics.

I. S. Paul Nation. 2006. How Large a Vocabulary Is Needed For Reading and Listening? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 63(1):59–82.

Charles Kay Ogden. 1968. *Basic English: international second language*. Harcourt, Brace & World.

Gustavo Paetzold and Lucia Specia. 2016a. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679.

Gustavo Paetzold and Lucia Specia. 2016b. 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, Portoroz, Slovenia. European Language Resources Association (ELRA).

Gustavo Paetzold and Lucia Specia. 2016c. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016d. 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, San Diego, California. Association for Computational Linguistics.

Francesco Ronzano, Ahmed Abura'ed, Luis Espinosa Anke, and Horacio Saggion. 2016. Taln at semeval-2016 task 11: Modelling complex words by contextual, lexical and semantic features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1011–1016, San Diego, California. Association for Computational Linguistics.

Matthew Shardlow. 2013a. 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, Sofia, Bulgaria. Association for Computational Linguistics (ACL).

Matthew Shardlow. 2013b. The cw corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics (ACL).

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)*, Vienna.

Edward L Thorndike and Irving Lorge. 1944. The teacher's wordbook of 30,000 words. new york: Columbia university, teachers college.

Michael P Toglia and William F Battig. 1978. *Handbook of semantic word norms.* Lawrence Erlbaum.

Michael Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, pages 6–11.

Krzysztof Wróbel. 2016. Plujagh at semeval-2016 task 11: Simple system for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 953–957, San Diego, California. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics (TACL)*, 3:283–297.

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*, New Orleans, United States. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017a. CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

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

Marcos Zampieri, Liling Tan, and Josef van Genabith. 2016. Macsaar at semeval-2016 task 11: Zipfian and character features for complexword identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1001–1005, San Diego, California. Association for Computational Linguistics.

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.

Qing Zeng-Treitler, Sergey Goryachev, Tony Tse, Alla Keselman, and Aziz Boxwala. 2008. Estimating Consumer Familiarity with Health Terminology: A Context-based Approach. *Journal of the American Medical Informatics Association: JAMIA*, 15(3):349–356.