Elnaz Davoodi, Leila Kosseim and Matthew Mongrain (2017). On the Influence of Contextual Features for the Identification of Complex Words, International Journal of Semantic Computing. pages 497-512, Volume 11, Issue 04, December 2017.

International Journal of Semantic Computing


# On the Influence of Contextual Features for the Identification of Complex Words

ELNAZ DAVOODI
*Dept. of Computer Science and Software Engineering*
*Concordia University*
*1515 Ste-Catherine Street West*
*Montréal, Québec, Canada H3G 2W1*
*e_davoo@encs.concordia.ca*

LEILA KOSSEIM
*Dept. of Computer Science and Software Engineering*
*Concordia University*
*1515 Ste-Catherine Street West*
*Montréal, Québec, Canada H3G 2W1*
*leila.kosseim@concordia.ca*

MATTHEW MONGRAIN
*Dept. of Computer Science and Software Engineering*
*Concordia University*
*1515 Ste-Catherine Street West*
*Montréal, Québec, Canada H3G 2W1*
*mmongrain@encs.concordia.ca*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

This paper evaluates the effect of the context of a target word on the identification of complex words in natural language texts. The approach automatically tags words as either complex or not, based on two sets of features: base features that only pertain to the target word, and contextual features that take the context of the target word into account. We experimented with several supervised machine learning models, and trained and tested the approach with the 2016 SemEval Word Complexity Data Set. Results show that when discriminating base features are used, the words around the target word can supplement those features and improve the recognition of complex words.

*Keywords*: Natural Language Processing; Text Simplification; Complex Word Identification; Feature Selection

## 1. Introduction

In order to map the semantics of natural language texts into a machine-readable format, it is important to understand when two textual elements (such as paragraphs,

sentences, phrases or even words) have the same meaning. Today, with the accessibility of the Web to a larger audience, many documents share the same meaning, but have different levels of readability. The articles on Simple English Wikipedia, for example, contain the same information as their counterparts on English Wikipedia, but their language is made simpler. In this context, being able to identify complex words, for instance to simplify documents for the sake of English language learners, becomes important.

In this paper, we evaluate the usefulness of contextual information for the identification of complex words in natural language texts. The approach automatically tags words as either complex or not, based on two sets of features: base features that only pertain to the target word, and contextual features that take the local context of the target word into account. We experimented with several supervised machine learning models and feature sets, and trained and tested the approach with the SemEval-2016 dataset. We show that when discriminating base features are used, words around the target word can provide important clues to improve the recognition of complex words.

This paper is organized as follows: Section 2 discusses previous work and situates our work within the field; Section 3 describes the SemEval 2016 Word Complexity dataset that we used; Section 4 explains the overall approach we followed to measure the influence of the context; Sections 5 and 6 describe and analyse a series of experiments using context to augment a variety of base features; finally, Section 7 presents future avenues of research.

## 2. Previous Work

Much research on text simplification has addressed the issue of lexical simplification (e.g. [1, 2]), a process that consists of identifying complex words and substituting them with simpler ones. Lexical simplification remains a challenge, as its first step, complex word identification, is still not performed accurately. The recent SemEval 2016 Complex Word Identification task [3] was a step in that direction. Given a set of sentences and a target word, systems had to automatically label the target word as being *complex* or *simple*. To do this, most approaches used standard supervised machine learning techniques (such as decision trees, nearest neighbors, SVM, or conditional random fields) and a variety of linguistic features. For example, [4], who achieved the highest F-score in the shared task, experimented with a variety of features such as the term and document frequency of the target word in the Simple English Wikipedia corpus [5], the length of sentences and words, the position of the target word within sentences, and word embedding. His experiments showed a minimal improvement when using many features, and thus a single feature was used for the task: the document frequency of the target word in Simple English Wikipedia. With this single feature, [4] achieved the highest F-score of 0.353 (the median was 0.171). However, since the training and test sets used in the shared task

come from the Simple English Wikipedia corpus themselves, it is not clear how well the approach would perform on a corpus from a different source. As a result, the state of the art performance leaves much room for improvement.

To our knowledge, no previous work has specifically investigated the effect of using the local context of a target word for complex word identification. On the other hand, the local context has long been used in many natural language applications such as word sense disambiguation (see [6] for a survey). The words surrounding a polysemous word have been shown to be particularly informative. As a result, a variety of strategies to consider the local context have emerged: using only open class words, using variable-sized windows, or using a decaying function to assign a weight to contextual features as a function of their distance from the target word (e.g. [7]). Inspired by work in the field of word sense disambiguation, we have evaluated the effect of the local context of a target word for the task of complex word identification. As described below, we have restricted our experiments to a fixed-size context and a uniform weighting scheme for all features.

## 3. Dataset

In order to develop and evaluate our approach, we used the SemEval 2016 Complex Word Identification Dataset[a] [3]. Each item in the dataset is composed of a sentence, a target word within that sentence, and a label indicating whether the word is *simple* or *complex*. For example, given the training instance of Example (1) below, the target word *explosion* is classified as simple; however in Example (2), the target *anoxic* is classified as complex.

| | | |
|---|---|---|
| (Example 1) | *During the attack, a blast from the explosion of gunpowder stored in Captain Jacobs's house was heard in Pittsburgh, 44 miles away.* | simple |
| (Example 2) | *Although anoxic events have not happened for millions of years, the geological record shows that they happened many times in the past.* | complex |

The SemEval-2016 dataset set does contain some peculiarities. First, the training set is much smaller than the test set, with 2,247 instances for training and 88,221 instances for testing. Second, both data sets are imbalanced, but do not have the same proportion of complex words, with 31% of words tagged as complex in the training set and only 5% tagged as complex in the test set. Statistics on the dataset are summarized in Table 1. Despite these characteristics, we used this dataset for our experiments, as it constitutes the largest complexity-tagged corpus assembled to date. However, in order to produce more reliable results, we re-balanced the training

[a]available at `http://alt.qcri.org/semeval2016/task11/index.php?id=results`

| **Data Set** | **# Instances** | **# Complex** | **% Complex** | **# Words per instance** |
|---|---|---|---|---|
| Training Set | 2,237 | 706 | 31% | 26.8 |
| Test Set | 88,221 | 4131 | 4.68% | 24.3 |

Table 1. Statistics of the SemEval Complex Word Identification Dataset

and testing sets to a more standard 80/20 split. Specifically, we concatenated the training and test sets as a single corpus, and used a random 80% static split for training, and the remaining 20% for testing.

## 4. Overall Approach

To measure the influence of the local context of a word for complex word identification, we experimented with several classifiers and a variety of feature sets. In total, 1619 experiments were conducted. In this paper, we only report the most interesting ones in Section 5.

### 4.1. *Classifiers*

We experimented with five different supervised machine learning models, as implemented in `scikit-learn` [8].

(1) A random forest model;
(2) An extremely randomized trees model;
(3) A Naïve Bayes model;
(4) A K-nearest neighbors model (with $k = 3$);
(5) A voting classifier implementing a weighted combination of the above four classifiers' predictions.

### 4.2. *Base and Contextual Features*

The approach uses two types of features: base features and contextual features.

**Base features** only consider information inherent to a word without looking at its context. For example, a word's length and frequency are the same regardless of its context. A variety of base features were used in this work, from simple word frequencies to psycholinguistic features. Table 2 describes these features briefly; while Section 5 describes each feature in detail.

Base features allow for the identification of target words out of context; however, the same word, when used in a particular sentence, may be perceived differently than it would be in another context. For example, in the training set, the word *happened* was perceived as being simple in Example (3), below, but as complex in Example (4).

| # | Feature | Description |
|---|---|---|
| 1. | goo_page_count | Google n-gram page count |
| 2. | goo_volume_count | Google n-gram volume count |
| 3. | goo_match_count | Google n-gram word frequency count |
| 4. | len | Word length |
| 5. | pos | POS tag |
| 6. | wd_syn | WordNet synonyms |
| 7. | mrc_conc | Concreteness level of the word |
| 8. | mrc_imag | Imagery level of the word |
| 9. | mrc_fam | Familiarity level of the word |
| 10. | mrc_aoa | Age of acquisition of the word |
| 11. | mrc_k_f_freq | Word frequency according to [9] |
| 12. | mrc_t_l_freq | Word frequency according to [10] |
| 13. | mrc_k_f_nsamp | Number of samples in which the word was found according to [9] |
| 14. | mrc_nphon | Number of phonemes of the word |
| 15. | mrc_brown_freq | Word frequency according to [11] |
| 16. | mrc_nsyl | Number of syllables of the word |
| 17. | mrc_k_f_ncats | Number of categories of text in which the word was found according to [9] |
| 18. | mrc_wtype | Specific syntactic category of the word |
| 19. | mrc_status | Status of the word according to [12] (eg. dialect, archaic, ...) |
| 20. | mrc_irreg | Plurality of the word |
| 21. | mrc_pdwtype | Coarse-grained syntactic category of the word |
| 22. | mrc_meanc | Meaningfulness rating of the word according to [13] |
| 23. | mrc_tq2 | Derivational variant of another word |
| 24. | mrc_meanp | Meaningfulness rating of the word according to Paivio [14] |
| 25. | mrc_var | Words which have the same spelling but different pronunciation and syntactic classes |
| 26. | mrc_cap | Whether or not the word is normally written with an initial capital letter |
| 27. | mrc_phon | Phonetic information of the word |
| 28. | mrc_alphsyl | Word is an abbreviation, suffix, prefix, is hyphenated or is a multi-word phrasal unit? |

Table 2. Description of the Base Features Used.

In the training set, this phenomenon occurred 94 times (4.18% of the corpus), and 304 times (1.38%) in the test set. To account for this, in addition to individual word features, we also took the local context of the target word into account.

(Example 3) *There are several stories about Mozart's final illness and death, and it is not easy to be sure what happened.* simple

| # | Feature | Used in . . . feature set | | | | |
|---|---|---|---|---|---|---|
| | | **SemEval (Section 5.1)** | **SemEval+Google (Section 5.2)** | **SemEval+Psycho (Section 5.3)** | **Single (Section 5.4)** | **Extended (Section 5.5)** |
| 1. | goo_page_count | | ✓ | | ✓ | ✓ |
| 2. | goo_volume_count | | ✓ | | | ✓ |
| 3. | goo_match_count | ✓ | ✓ | ✓ | | ✓ |
| 4. | len | ✓ | ✓ | ✓ | | ✓ |
| 5. | pos | ✓ | ✓ | ✓ | | ✓ |
| 6. | wd_syn | ✓ | ✓ | ✓ | | ✓ |
| 7. | mrc_conc | | ✓ | ✓ | | ✓ |
| 8. | mrc_imag | | | ✓ | | ✓ |
| 9. | mrc_fam | | | ✓ | | ✓ |
| 10. | mrc_aoa | | | ✓ | | ✓ |
| 11. | mrc_k_f_freq | | | | | ✓ |
| 12. | mrc_t_l_freq | | | | | ✓ |
| 13. | mrc_k_f_nsamp | | | | | ✓ |
| 14. | mrc_nphon | | | | | ✓ |
| 15. | mrc_brown_freq | | | | | ✓ |
| 16. | mrc_nsyl | | | | | ✓ |
| 17. | mrc_k_f_ncats | | | | | ✓ |
| 18. | mrc_wtype | | | | | ✓ |
| 19. | mrc_status | | | | | ✓ |
| 20. | mrc_irreg | | | | | ✓ |
| 21. | mrc_pdwtype | | | | | ✓ |
| 22. | mrc_meanc | | | | | ✓ |
| 23. | mrc_tq2 | | | | | ✓ |
| 24. | mrc_meanp | | | | | ✓ |
| 25. | mrc_var | | | | | ✓ |
| 26. | mrc_cap | | | | | ✓ |
| 27. | mrc_phon | | | | | ✓ |
| 28. | mrc_alphsyl | | | | | ✓ |

Table 3. Features Used in Each Experiment.

(Example 4) *Although anoxic events have not happened for millions of years, the geological record shows that they happened many times in the past.* complex

**Contextual features** consider the surrounding words in order to classify the target word. Hence the same target word may be classified as *complex* in a sentence, but *simple* in another based on the words around it.

Contextual features supplement the base features of the target word with the same features of its surrounding words, in addition to using word sequence information. In all experiments reported in Section 5, we experimented with seven different context sizes varying from $c = 0$ to $c = 6$. For each context size $c$, we took into account:

the $n$ base features of the target word
+ (at most) the $n$ base features of the previous $c$ words
+ (at most) the $n$ base features of the following $c$ words
+ two word sequence features

This gave rise to a maximum of $n + (2 \times c \times n) + 2$ features. Note that when

$c = 0$, only the base features of the target word are considered and no context is taken into account. In addition, because the instances in the data set are based on sentences, we do not cross sentence boundaries to extract context, hence explaining the *at most* above. For example, if the target word is the $4^{th}$ word of a sentence of 8 words, then with a context size of $c = 5$ only the features of the target word, those of the 3 previous words, and those of the following 4 words are considered.

The last two features (word sequence features) take into account the probability of seeing a particular word-based n-gram of size $c$ in the context of a complex word compared to the probability of seeing the same n-gram in the context of a simple word. To do this, we used the SemEval training set (see Section 3) to build a language model for complex-word contexts and a language model for simple-word contexts, and used the probability of the n-gram occurring in the context of the target word for each model (complex-word and simple word) as additional features.

## 5. Experiments

As indicated in Section 4, five supervised learning models were used with a variety of features. In this section, we describe the results of five different experiments to evaluate the effect of contextual features. A summary of the features used in each experiment is available in Table 3.

### 5.1. *SemEval Features*

In this first experiment, we used the four features proposed in [15] at the SemEval shared task. This includes four linguistic features and one psycholinguistic feature. Much research has linked the comprehension of words to their psycholinguistic features (e.g. [16, 17]). To take this information into account, we used the Medical Research Council (MRC) psycholinguistic database [14]. This electronic resource contains 150,837 words annotated with a score for up to 26 linguistic and psycholinguistic features collated from the variety of other sources. We did not use the MRC for its linguistic features (such as frequency or syntactic category) as we had already used more modern resources for these. The five features used are described below.

**1) Frequency of the target word (`goo_match_count`)** A great deal of work in linguistics and psycholinguistics has highlighted the relationship between the frequency of linguistic elements (such as words, expressions and grammatical structures) within a text and their level of complexity (e.g. [18, 19, 20]). In light of these observations, our first feature takes into account the frequency of the target word in English. For this, we used the of Google Web1T N-gram corpus [21]. The Google corpus is a collection of English one- to five-grams tagged with their frequencies, organized by year, which was mined from approximately 1 trillion words from the web. In order to focus only on recent word usages and to reduce the size of the corpus, we only considered the frequency of the target word (i.e. Google's `goo_match_count`

value) in sources indexed after year 2000. This way, we reduced the influence of once frequently used but now obsolete words.

**2) Part of speech tag of the target word (`pos`)** A word may be assigned different parts of speech depending on its use in context. For example, the word *happening* may be used as a verb (in gerund form), as a noun or as an adjective. To consider that all usages of the same word have the same complexity level would be a generalisation. For example, a word used as a verb may be perceived as complex, whereas its use as a noun may not. To account for this, we parsed each sentence of the dataset with the Stanford POS tagger [22] and used the part of speech tag of the target word as a feature.

**3) Number of synonyms of the target word (`wd_syn`)** Our analysis of the training set revealed that complex words tend to have fewer synonyms than simpler words. To determine this, we used WordNet [23] to compute the number of synonyms of each word tagged as complex in the training set versus the number of synonyms of words tagged as simple. Results show that 33.65% of complex words but only 24.10% of simple words have fewer than four synonyms. Given this observation, we considered the number of synonyms of the target word as one of our features.

**4) Length of the target word (`len`)** Based on the work of traditional text complexity measures such as the Flesch index [24], we took into account the length of a word, in terms of the number of characters it contains, as a feature to determine its complexity.

**5) The concreteness level of the target word (`mrc_conc`)** Research in psycholinguistics has linked word recognition and comprehension to the use of more concrete words versus more abstract words (e.g. [17]). For example, words that refer to objects, materials, or persons are more concrete and hence easier to comprehend. To take this information into account, we used the *concreteness* measure of the MRC. This concreteness measure is available for 8,228 words and is indicated by an integer value ranging from 100 (very abstract) to 700 (very concrete). For this feature, if the target word has a concreteness value in the MRC, we use its value. If a word has no concreteness value in the MRC, we assign it a value of 400 (average).

Table 4 shows the precision, recall and F-score of the Naïve Bayes and Random Forest classifiers for the complex words computed with the official evaluation script of the SemEval 2016 shared task [3]. We only report here the results of the Naïve Bayes classifier used as a baseline, and the best performing classifier: a Random Forest model. On average, accuracy is around 0.85, but given the imbalanced dataset, this is not an informative measure, and is therefore not reported in Table 4. Note also that the precision, recall and F-score are given in terms of the complex class, which is the minority class in our data set and therefore the more

| **Feature Set Size ($n$)** | **Classifier** | **Context Size ($c$)** | **Recall** | **Precision** | **F-score** |
|---|---|---|---|---|---|
| SemEval (5) | NB | 6 | 0.070 | 0.891 | 0.129 |
| SemEval (5) | NB | 0 | 0.055 | 0.970 | 0.104 |
| SemEval (5) | NB | 1 | 0.063 | 0.940 | 0.117 |
| SemEval (5) | NB | 2 | 0.067 | 0.917 | 0.125 |
| SemEval (5) | NB | 3 | 0.067 | 0.915 | 0.125 |
| SemEval (5) | NB | 4 | 0.067 | 0.916 | 0.125 |
| SemEval (5) | NB | 5 | 0.069 | 0.897 | 0.129 |
| SemEval (5) | RF | 0 | 0.146 | 0.699 | 0.241 |
| SemEval (5) | RF | 1 | 0.165 | 0.687 | 0.267 |
| SemEval (5) | RF | 2 | 0.174 | 0.683 | 0.277 |
| SemEval (5) | RF | 3 | 0.174 | 0.681 | 0.277 |
| SemEval (5) | RF | 4 | 0.175 | 0.680 | 0.279 |
| SemEval (5) | RF | 5 | 0.179 | 0.669 | 0.282 |
| SemEval (5) | RF | 6 | 0.178 | 0.667 | 0.281 |

Table 4. Performance of the Naïve Bayes (NB) and Random Forest (RF) Models with the SemEval Feature Set.

difficult class to identify. As the table shows, the F-score of the Naïve Bayes (around 0.12) is typically much lower than that of the Random Forest (around 0.27). The overall performance of our system is in line with that of the other participants at the 2016 SemEval Complex Word Identification task (e.g. [25]). As Table 4 shows, the Random Forest model compares favorably with the median F-score of 0.171 at SemEval, but is still far from the best score of 0.353 [4]. In addition, the Random Forest model performs significantly better when the local context of target words is taken into account. The classifier gradually increases its F-score as more contextual features are used, reaching 0.281 with $c = 6$ from 0.241 with $c = 0$.

### 5.2. *SemEval+Google Features*

In order to verify the effect of the context with a different feature set, we ran the same experiment again, but augmented the SemEval features (see Section 5.1) with all frequencies provided in the Google Ngram Viewer Version 1. This includes:

(1) `goo_page_count`: number of Google pages containing the word
(2) `goo_volume_count`: number of Google books (or volumes) containing the word
(3) `goo_match_count`: overall frequency of the word.

As indicated in Table 3, this gave rise to seven features.

The same five classifiers were used and, again, the Random Forest classifier achieved the best performance. The results, shown in Table 5, show that the use of the additional Google frequencies do not improve the F-score. With no contextual window, the F-score of 0.216 is lower than the F-score of 0.241 reached by the five SemEval features alone. This is surprising, as [4] identifies document frequency as very informative for the task. However, when examining the influence of context, surrounding words again seem to provide useful clues, as the F-score increases steadily from 0.216 to 0.267 as the size of the contextual window increases from $c = 0$ to $c = 6$.

### 5.3. *SemEval+Psycho Features*

In [15], a feature selection process identified the MRC's psycholinguistic features as very discriminating. Hence, as a third experiment, we augmented the five SemEval features with three more psycholinguistic features, resulting in a total of eight features (see Table 3).

**1) The imagery level of the target word (`mrc_imag`)** Words with a high level of imagery evoke a strong sensory experience or arouse mental images quickly and easily and are therefore more likely to be recalled [26]. To account for these, we used the MRC's *imagery* score. This feature is indicated for 4,825 words on a scale of 100 to 700. For example, the word *accident* has a value of 518, whereas the word *after* has a lower value of 217. As with the *concreteness* level, if a word has no *imagery* value in the MRC, we assign it the average value of 400.

**2) The familiarity level of the target word (`mrc_fam`)** Likewise, we used the *familiarity* level, which is indicated for 4,920 words in the MRC. Scores range from 100 to 700, where higher scores indicate greater familiarity. For example, the word *adze* has a low familiarity score, whereas *eating* has a high score. As with the other features, if a word has no *familiarity* value in the MRC, we assign it a value of 400.

**3) The age of acquisition of the target word (`mrc_aoa`)** *Age of acquisition* is an indication of the age when a word is typically learned and has been shown to be correlated to memory processes (e.g. [27]). The MRC indicates this feature for 3,503 words, again on a scale of 100 (early learning) to 700 (late learning). As with the other psycholinguistic features, if a word has no *age of acquisition* value in the MRC, we assign it a value of 400.

As Table 6 shows, although these features produced the lowest F-score compared to the previous two experiments, the Random Forest model again performs significantly better when the local context is taken into account. The classifier gradually increases its F-score as more contextual features are used, and peaks at $c = 5$.

| **Feature Set Size ($n$)** | **Classifier** | **Context Size ($c$)** | **Recall** | **Precision** | **F-score** |
|---|---|---|---|---|---|
| SemEval+Google (7) | RF | 0 | 0.129 | 0.653 | 0.216 |
| SemEval+Google (7) | RF | 1 | 0.147 | 0.668 | 0.241 |
| SemEval+Google (7) | RF | 2 | 0.157 | 0.670 | 0.255 |
| SemEval+Google (7) | RF | 3 | 0.159 | 0.673 | 0.258 |
| SemEval+Google (7) | RF | 4 | 0.162 | 0.680 | 0.262 |
| SemEval+Google (7) | RF | 5 | 0.164 | 0.691 | 0.266 |
| SemEval+Google (7) | RF | 6 | 0.165 | 0.696 | 0.267 |

Table 5. Performance of the Random Forest (RF) Model with the SemEval+Google Feature Set.

| **Feature Set Size ($n$)** | **Classifier** | **Context Size ($c$)** | **Recall** | **Precision** | **F-score** |
|---|---|---|---|---|---|
| SemEval+Psycho (8) | RF | 0 | 0.501 | 0.112 | 0.184 |
| SemEval+Psycho (8) | RF | 1 | 0.391 | 0.144 | 0.211 |
| SemEval+Psycho (8) | RF | 2 | 0.340 | 0.140 | 0.199 |
| SemEval+Psycho (8) | RF | 3 | 0.537 | 0.149 | 0.233 |
| SemEval+Psycho (8) | RF | 4 | 0.469 | 0.176 | 0.256 |
| SemEval+Psycho (8) | RF | 5 | 0.548 | 0.170 | 0.260 |
| SemEval+Psycho (8) | RF | 6 | 0.592 | 0.107 | 0.181 |

Table 6. Performance of the Random Forest (RF) Model with the SemEval+Psycho Feature Set.

However, unlike previous experiments, the F-score decreases when $c = 6$. It is not clear what causes this drop, but we suspect that the effect of the context is not as strong for this set of features, simply because the base features themselves are not as discriminating as the SemEval or the SemEval+Google feature sets.

### 5.4. *Single Feature*

Because the addition of new features did not seem to improve the overall F-score, in this fourth experiment, we meant to evaluate the influence of the context on a much smaller set of highly discriminating features. [4] identified the document frequency as the most discriminating feature at SemEval 2016; therefore, we experimented with the use of this feature on its own. In a sense, this experiment is similar in spirit to evaluating the influence of contextual features on the best performing system at SemEval 2016 [4]. However, recall that the SemEval dataset was created from the Simple Wikipedia corpus. As opposed to [4], who used frequencies from the same corpus, we used frequency counts from Google in order to introduce no bias towards that particular corpus and hence avoiding any overfitting. Therefore, instead of using Wikipedia's document frequency, we used Google's page count

| **Feature Set Size ($n$)** | **Classifier** | **Context Size ($c$)** | **Recall** | **Precision** | **F-score** |
|---|---|---|---|---|---|
| Single (1) | RF | 0 | 0.060 | 0.322 | 0.101 |
| Single (1) | RF | 1 | 0.062 | 0.184 | 0.093 |
| Single (1) | RF | 2 | 0.043 | 0.142 | 0.066 |
| Single (1) | RF | 3 | 0.047 | 0.158 | 0.072 |
| Single (1) | RF | 4 | 0.047 | 0.142 | 0.070 |
| Single (1) | RF | 5 | 0.054 | 0.167 | 0.082 |
| Single (1) | RF | 6 | 0.063 | 0.185 | 0.094 |

Table 7. Performance of the Random Forest (RF) Model with the Single Feature (`goo_page_count`).

frequency (`goo_page_count`).

Table 7 shows the results of the Random Forest model using only Google's page count frequency feature. The results show that the use of this single feature is not sufficient for the task of complex word identification, as its highest F-score is 0.101. This is very far from the result of 0.353 reported in [4], and we suspect that this may be due in part to an overfitting of their approach to the corpus. Table 7 also shows that the context does not seem to help in improving the performance as the F-score does not increase with the context. Clearly, if weakly discriminating base features are used, the same features of surrounding words cannot help.

### 5.5. *Extended Feature Set*

As a last experiment, we used the full set of 28 features described in Table 2. This includes:

(1) All three frequencies provided in the Google Ngram Viewer Version 1.
(2) 22 of the 26 features[b] of the MRC psycholinguistic database [14].

The results, reported in Table 8, show a significant improvement of the F-score compared to the use of the SemEval+Psycho and SemEval+Google features (see Sections 5.3 and 5.2) with a consistent average around 0.28 compared to 0.22. The use of all features achieves the same F-score of the orginal SemEval features (see Section 5.1). In addition, similarly to the results of Table 4, the performance of the extended feature set shows an increase in F-score as more context is considered.

## 6. Analysis

Figure 1 shows the F-score of all five experiments described in Section 5 graphically. As the figure shows, all feature sets benefit from the use of contextual information

[b]We did not include: `dphon` (phonetic transcription) and `stress` (stress pattern) which are phonetic in nature, `nlet` (number of letters in the word) which is already taken into account, and `word` (the word itself) which is too sparse and therefore uninformative.

| **Feature Set** | **Classifier** | **Context Size ($c$)** | **Recall** | **Precision** | **F-score** |
|---|---|---|---|---|---|
| Extended (28) | RF | 0 | 0.145 | 0.699 | 0.241 |
| Extended (28) | RF | 1 | 0.165 | 0.687 | 0.266 |
| Extended (28) | RF | 2 | 0.174 | 0.683 | 0.277 |
| Extended (28) | RF | 3 | 0.174 | 0.681 | 0.277 |
| Extended (28) | RF | 4 | 0.175 | 0.680 | 0.279 |
| Extended (28) | RF | 5 | 0.179 | 0.669 | 0.282 |
| Extended (28) | RF | 6 | 0.178 | 0.667 | 0.281 |

Table 8. Performance of the Random Forest (RF) Model with the Extended Feature Set

when using a Random Forest classifier except when using only the Google page count feature.

It is interesting to note that the influence of context on the models constructed is similar regardless of whether the SemEval features, the extended set of features, or the SemEval+Google feature set is used (see the top 3 lines in Figure 1). However, the F-score of the model built based on the SemEval+Google feature set is lower than that of the other two models with the same contextual window size. This seems to show that various word frequencies are not sufficient for complex word identification. In addition, we suspect that the erratic behavior of the Psycho+SemEval features is due to the limited coverage of the MRC dictionary. Further research would be necessary to analyze this effect fully.

A closer look at contextual window sizes shows that the top performing models (SemEval, Extended set, and SemEval+Google) benefit mostly from an increase in window size from $c = 0$ to $c = 2$. The increase in F-score is not as strong for $c > 3$ which is to be expected as words further from the target word have a smaller semantic influence on the target word. This conclusion is in line with work in word sense disambiguation (e.g. [7]) where the sense of a target word is very dependent on its local context, but words at longer distance have little influence.

Overall, the Random Forest models using the extended feature set, made of 28 features, and the SemEval feature set, made of five features, achieve the best overall F-score. The original SemEval feature set with a contextual window size of $c = 6$ therefore constitutes the preferred feature set as it achieves the best performance, yet requires fewer resources than the extended feature set.

## 7. Conclusion and Future Work

In this paper we have evaluated the influence of context in the identification of complex words in natural language texts. We have described five experiments using a Random Forest classifier with different feature sets evaluated on the SemEval 2016 Word Complexity Data Set [3]. Results show that when using strongly discriminating base features, the words around the target word can provide important clues

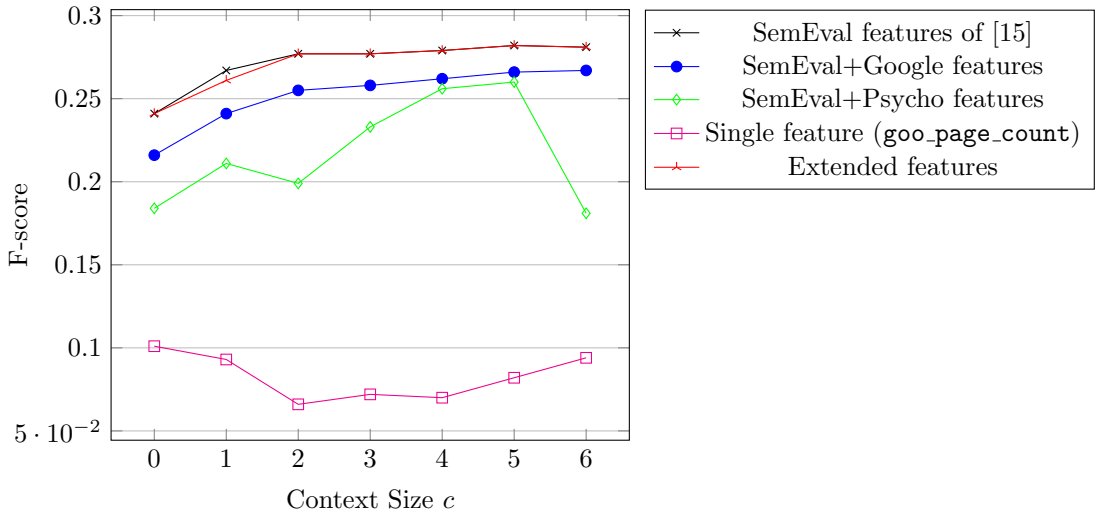


Fig. 1. F-score of the Random Forest Models with the Various Feature Set as a Function of the Context Size

that improve the recognition of complex words. However, when weakly discriminating features, such as Google page count, are used, contextual information is not useful.

Further investigation could explore the feature weights learned by the Random Forest classifiers, in order to further reduce the feature sets with minimal loss of predictive performance.

Another interesting line of further research is the issue of the unbalanced dataset. As shown in Section 3, the dataset does not contain many training instances for complex words. Using under-sampling or over-sampling approaches, as in [28], might lead to better performance. Finally, in our experiments, we used windows of fixed size, considered all the words in the window and assigned all features the same weight. It would be interesting to see if using variably-sized contexts, filtering words within the context, or using different weights for contextual features based on their distance from the target word would lead to better results.

**Acknowledgement**

The authors would like to thank the anonymous reviewers for their feedback on an earlier version of the paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

## References

[1] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait, "Practical simplification of English newspaper text to assist aphasic readers," in *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, Wisconsin, Jul. 1998, p. 7–10.

[2] O. Biran, S. Brody, and N. Elhadad, "Putting it simply: A context-aware approach to lexical simplification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*, Portland, Jul. 2011, p. 496–501.

[3] G. H. Paetzold and L. Specia, "SemEval 2016 Task 11: Complex Word Identification," in *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, Jun. 2016, pp. 560–569.

[4] K. Wròbel, "PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification," in *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, Jun. 2016, pp. 953–957.

[5] W. Coster and D. Kauchak, "Simple English Wikipedia: A New Text Simplification Task," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT-2011) Short Papers - Volume 2*, 2011, p. 665–669.

[6] N. Ide and J. Véronis, "Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art," *Computational Linguistics*, vol. 24, no. 1, p. 2–40, Mar. 1998.

[7] B. Brosseau-Villeneuve, N. Kando, and J.-Y. Nie, "Construction of context models for word sense disambiguation," *Journal of Natural Language Processing*, vol. 18, no. 3, p. 217–245, 2010.

[8] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, A. G. Peter Prettenhofer, J. V. Jaques Grobler, Robert Layton, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, Pragues, Czech Republic, Sep. 2013.

[9] H. Kucera and W. Francis, *Computational Analysis of Present-Day American English*. Providence: Brown University Press, 1967.

[10] E. Thorndike and I. Lorge, *The teacher's word book of 30,000 words*. Columbia University: New York: Teachers College, 1944.

[11] G. D. A. Brown, "A frequency count of 190,000 words in the London-Lund Corpus of English Conversation," *Behavior Research Methods, Instruments, & Computers*, vol. 16, no. 6, pp. 502–532, 1984.

[12] J. Dolby, H. Resnikoff, and F. MacMurray, "A tape dictionary for linguistic experiments," in *Proceedings of the American Federation of information processing societies:Fall Joint Computer Conference, Volume 24*, Baltimore, MD, 1963, p. 419–423.

[13] M. Toglia and W. Battig, *Handbook of semantic word norms*. Hillsdale, N.J. : Lawrence Erlbaum Associates, 1978.

[14] M. Wilson, "MRC psycholinguistic database: Machine-usable dictionary, version 2.00," *Behavior Research Methods, Instruments, & Computers*, vol. 20, no. 1, p. 6–10, 1988.

[15] E. Davoodi and L. Kosseim, "CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic Features for Complex Word Identification," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, Jun. 2016, pp. 982–985.

[16] F. R. Vellutino and D. M. Scanlon, "Free recall of concrete and abstract words in poor and normal readers," *Journal of Experimental Child Psychology*, vol. 39, no. 2, pp. 363–380, Apr. 1985.

[17] N. Gee, D. Nelson, and D. Krawczyk, "Is the concreteness effect a result of underlying network interconnectivity?" *Journal of Memory and Language*, vol. 40, pp. 479–947, 1999.

[18] D. Roland, F. Dick, and J. L. Elman, "Frequency of basic English grammatical structures: A corpus analysis," *Journal of Memory and Language*, vol. 57, no. 3, p. 348–379, 2007.

[19] E. Bates, S. D'Amico, T. Jacobsen, A. Székely, E. Andonova, A. Devescovi, D. Herron, C. C. Lu, T. Pechmann, C. Pléh, N. Wicha, K. Federmeier, I. Gerdjikova, G. Gutierrez, D. Hung, J. Hsu, G. Iyer, K. Kohnert, T. Mehotcheva, A. Orozco-Figueroa, A. Tzeng, and O. Tzeng, "Timed picture naming in seven languages," *Psychonomic Bulletin Review*, vol. 10, no. 2, p. 344–380, 2003.

[20] J. G. Snodgrass and M. Vanderwart, "A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity," *Journal of Experimental psyshology: Human learning and memory*, vol. 6, no. 2, p. 174–215, 1980.

[21] J.-B. Michel *et al.*, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, no. 6014, p. 176–182, 2011.

[22] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL)*, Edmonton, May 2003, p. 173–180.

[23] G. A. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, vol. 38, no. 11, p. 39–41, Nov. 1995.

[24] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel," DTIC Document, Tech. Rep., 1975.

[25] F. Ronzano, A. Abura'ed, L. Espinosa Anke, and H. Saggion, "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)*, San Diego, California, June 2016, pp. 1011–1016.

[26] N. Ellis and A. Beaton, "Psycholinguistic determinants of foreign language vocabulary acquisition," *Language Learning*, vol. 43, pp. 559–617, 1993.

[27] S. Roodenrys, C. Hulme, J. Alban, A. Ellis, and G. Brown, "Effects of word frequency and age of acquisition on short-term memory span," *Memory Cognition*, vol. 22, p. 695–701, 1994.

[28] H. Almeida, M.-J. Meurs, L. Kosseim, and A. Tsang, "Data Sampling and Supervised Learning for HIV Literature Screening," *IEEE Transactions on NanoBioscience*, vol. 15, no. 4, p. 354–361, 2016.