Elnaz Davoodi, Leila Kosseim and Matthew Mongrain (2017). A Context-Aware Approach for the Identification of Complex Words in Natural Language Texts. In Proceedings of the 11th IEEE International Conference on Semantic Computing (IEEE ICSC 2017). pages 97--100. Jan 30-Feb 1. San Diego, USA.

A Context-Aware Approach for the Identification of Complex Words in Natural Language Texts

Elnaz Davoodi, Leila Kosseim and Matthew Mongrain
Depart. of Computer Science and Software Engineering
Concordia University
1515 Ste-Catherine Street West
Montréal, Québec, Canada H3G 2W1
{e_davoo, kosseim, mmongrain}@encs.concordia.ca

***Abstract*—This paper evaluates the effect of the context 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 SemEval-2016 dataset. Results show that considering contextual features significantly improves the identification of complex words by reaching an F-measure of 0.260 compared to 0.184 without them.**

I. 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 (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 readability levels. 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 them for the sake of English language learners, becomes important.

In this paper, we describe an approach to 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 trained and tested the approach with the SemEval-2016 dataset. Results show that contextual features are just as important as features pertaining to the target word alone, and that considering them can significantly improve the recognition of complex words.

II. 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 Word Complexity 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-measure in the shared task, experimented with a variety of features such as the term and document frequency and 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 shared task: the document frequency of the target word in Simple English Wikipedia. With this single feature, [4] achieved the highest F-measure of 0.353 (the median was 0.171). However, since the data set used in the shared task comes from the Simple English Wikipedia corpus, it is not clear how the approach would behave given a corpus from a different source. As a result, the state of the art performance still leaves much room for improvement.

To our knowledge, no previous work has specifically investigated the effect of using the local context of the 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 very informative. As a result, many approaches have experimented with a variety of strategies to consider the local context: 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 word sense disambiguation, we have evaluated the effect of local context for the task of complex word identification. As described below, we have restricted our experiments to fixed-sized context and a uniform weight for all features.

III. Dataset

In order to develop and evaluate our approach, we used the SemEval 2016 Complex Word Identification Dataset[1] [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 (1) below, the target word *explosion* is classified as simple; however in (2), the target *anoxic* is classified as complex.

(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

(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 compared to only 5% in the test set. Despite these characteristics, we used this dataset, as it constitutes the largest complexity-tagged corpus to date. However, given the disproportionate size of the test set compared to the training set, we only used the first 22,000 instances of the test set for our experiments. Statistics on the dataset used are provided in Table I.

	# Instances	# Complex	% Complex	# Words per instance
Training	2,237	706	31%	26.8
Testing	22,000	1,167	5%	24.3

TABLE I
STATISTICS OF THE DATASET

IV. Approach

We experimented with various supervised machine learning models and two types of features.

a) Base features: only consider information inherent to a word to a word without looking at its context. For example, length, polysemy, or a word's frequency in the Google N-gram corpus do not vary based on the sentence a word appears in. These features are discussed in Section IV-A.

b) 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. These features are described in detail in Section IV-B.

A. Base Features

Base features consider the information that is inherent to target word only. We used an extension of the features proposed in [8] and considered 8 base features: 4 linguistic features and 4 psycholinguistic features.

1) Linguistic features include: a) the frequency of the target word, b) its part of speech tag, c) the number of synonyms the target word has, and d) its length.
2) Psycholinguistic features include: a) the *abstraction* level of the target word, b) its *imagery*, c) its *familiarity* level, and d) its *age of acquisition*.

Each feature is described below.

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

1) Linguistic features of the target word:

a) Frequency of the target word: 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. [9], [10], [11]). In light of these observations, our first feature takes into account the frequency of the target word in English. For this, we used the Google Web1T N-gram corpus [12]. 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 in sources indexed after year 2000. This way, we reduced the influence of once frequently used but now obsolete words.

b) Part of speech tag of the target word: 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 senses 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 [13] and used the part of speech tag of the target word as a feature.

c) Number of synonyms of the target word: Our analysis of the training set revealed that complex words tend to have fewer synonyms than simpler words. To determine this, we used WordNet [14] 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 have less than 4 synonyms; where as this number drops to 24.10% for simple words. Given this observation, we considered the number of synonyms of the target word as one of our features.

d) Length of the target word: Based on the work of traditional text complexity measures such as the Flesch index [15], 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.

2) Psycholinguistic features of the target word: 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 [18]. This electronic resource contains 150,837 words annotated with their score for up to 26 linguistic and psycholinguistic features. We did not use the

MRC for its linguistic features (such as frequency or syntactic category) as we already used more extensive resources for these types of features (see Section IV-A1). Instead, we used its psycholinguistic information, which includes scores for imagery, concreteness, familiarity, and age of acquisition.

a) The concreteness level of the target word: 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).

b) The imagery level of the target word: 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 [19]. 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 a value of 400.

c) The familiarity level of the target word: 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.

d) The age of acquisition of the target word: 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. [20]). 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.

B. Context-Aware Features

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

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

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

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. To do this, we augmented the eight base features of the target word of Section IV-A with the same eight features of its surrounding words, and used word sequence features. We experimented with six different window sizes varying from $n = 0$ to $n = 7$. For each window size n, we took into account:

the 8 base features of the target word
+ (at most) the 8 base features of the previous n words
+ (at most) the 8 base features of the following n words
+ two word sequence features

This gave rise to a maximum of $8 + 2 \times n \times 8$ + 2 features. Note that when $n = 0$, 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 the local context, hence explaining the *at most* above. For example, Hence if the target word is the 4^{th} word of a sentence of eight words, then with a window size of $n = 5$ only the features of the target word, those of the three previous words, and those of the following four words are considered.

The last two features (word sequence features) take into account the probability of seeing a particular word-based ngram of size n in the context of a complex word compared to the probability of seeing the ngram in the context of a simple word. To do this, we used the training set 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.

Classifier	Context Size (n)	Recall	Precision	F-measure
Naïve Bayes	0	0.745	0.090	**0.161**
Naïve Bayes	1	0.351	0.088	0.141
Naïve Bayes	2	0.268	0.104	0.150
Naïve Bayes	3	0.642	0.088	0.155
Naïve Bayes	4	0.351	0.103	0.159
Naïve Bayes	5	0.391	0.118	**0.181**
Naïve Bayes	6	0.001	0.024	0.003
Random Forest	0	0.501	0.112	**0.184**
Random Forest	1	0.391	0.144	0.211
Random Forest	2	0.340	0.140	0.199
Random Forest	3	0.537	0.149	0.233
Random Forest	4	0.469	0.176	0.256
Random Forest	5	0.548	0.170	**0.260**
Random Forest	6	0.592	0.107	0.181

TABLE II
PERFORMANCE OF THE LEARNING MODELS WITH THE TEST SET. RECALL, PRECISION AND F-SCORE ARE GIVEN IN TERMS OF THE COMPLEX CLASS – THE LEAST REPRESENTED CLASS.

V. RESULTS AND DISCUSSION

We experimented with various machine learning models trained on the features described above. As a baseline, we used

a Naïve Bayes classifier, and report on the best performing classifier: a Random Forest model. Table II shows the precision, recall and F-measure of the classifiers for the complex words in the test set, computed with the official evaluation script of the SemEval 2016 shared task [3].

Given the imbalanced dataset, accuracy is not an informative measure, and therefore is not given[2]. In addition, precision, recall and the F-measure are given in terms of the complex class, the minority class and the harder to identify.

As can be seen in Table II, both classifiers perform significantly better when a local context of 5 words is taken into account. Both classifier gradually increase their F-measure as more contextual features are used, and peak at $n = 5$. Using more than 5 words of local context lowers the F-measure significantly. In other words, a little context ($0 < n < 5$) is more useful than no context at all ($n = 0$), but beyond a certain point, further words have little 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 long distance words have little influence.

VI. Conclusion and Future Work

In this paper we have described the influence of context in the identification of complex words in natural language texts. Using the SemEval Word Complexity Data Set [3], we showed that the words around the target word are just as important to consider as the target word itself, and considering a small local context (5 in our experiments) can significantly improve the identification of complex words.

As future work, it would be interesting to see if the local context can be used to improve the current state of the art in complex word identification. [4] achieved the best F-measure at SemEval 2016, without using contextual information. Adding contextual features to their approach may improve their already high F-measure. Another interesting line of research is the issue of the unbalanced dataset. As shown in Section III, the dataset does not contain many training instances for complex words. Using under-sampling or over-sampling approaches, as in [21], might lead to a better performance. Finally, in our experiments, we used windows of fixed size, considered all the words in the window and all features were assigned the same weight. It would be interesting to see if using variable-size contexts, filtering words within the context or using different weights for contextual features based on their distance to the target word would lead to better results.

Acknowledgement

The authors would like to thank the anonymous reviewers for their feedback on the paper. This work was financially supported by NSERC.

[2]On average, accuracy is around 0.85.

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," *Compuutational 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] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

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

[19] E. N.C. and B. A., "Psycholinguistic determinants of foreign language vocabulary acquisition," *Language Learning*, vol. 43, pp. 559–617, 1993.

[20] 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.

[21] 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.