Deep Learning Architecture for Complex Word Identification

Dirk De Hertog
ITEC, imec, KU Leuven
dirk.dehertog@kuleuven.be

Anaïs Tack
CENTAL, Univ. catholique de Louvain
ITEC, imec, KU Leuven
F.R.S.-FNRS Research Fellow
anais.tack@uclouvain.be

Abstract

We describe a system for the CWI-task that includes information on 5 aspects of the (complex) lexical item, namely distributional information of the item itself, morphological structure, psychological measures, corpus-counts and topical information. We constructed a deep learning architecture that combines those features and apply it to the probabilistic and binary classification task for all English sets and Spanish. We achieved reasonable performance on all sets with best performances seen on the probabilistic task, particularly on the English news set (MAE 0.054 and F1-score of 0.872). An analysis of the results shows that reasonable performance can be achieved with a single architecture without any domain-specific tweaking of the parameter settings and that distributional features capture almost all of the information also found in hand-crafted features.

1 Introduction

In general, complex word identification (CWI) aims to identify words that are perceived as difficult for a given target audience. As such, children (De Belder and Moens, 2010), foreign language learners (Paetzold and Specia, 2016c) and readers suffering from aphasia (Devlin and Tait, 1998), dyslexia (Rello et al., 2013) or autism spectrum disorder (Štajner et al., 2017) will struggle with different words.

The goal of the current CWI shared task (Yimam et al., 2018) is to predict which words can be difficult for a non-native speaker, based on annotations collected from a mixture of native and non-native speakers. The instructions for the English dataset are formulated so that the annotator marks the words he thinks are problematic for children, non-native speakers, or people with language disabilities.

Having such a diverse target audience requires a system that includes a variety of information at different levels of linguistic description. We include information that covers 5 aspects of the lexical item at hand, namely distributional information of the item itself, morphological structure, psychological measures, corpus-counts and topical information. With the exception of the psychological measures, all can be readily trained by an appropriate neural network architecture and/or acquired from large-scale corpora.

We train a neural network to integrate said sources of information and apply it to the probabilistic and the binary complexity assessment for the three English datasets and the Spanish one.

2 Related Work

2.1 Complex Word Identification

The task of complex word identification has often been regarded as a critical first step for automatic lexical simplification (Shardlow, 2014). Indeed, erroneously identifying or failing to identify words as complex is likely to trigger important errors in the simplification pipeline. As a result, a growing number of studies have been dedicated specifically to complex word identification and have focused on developing accurate statistical learning methods and on collecting appropriate gold standards (Paetzold and Specia, 2016a; Yimam et al., 2017a,b; Štajner et al., 2017)

Complex word identification has only relatively recently been framed as a machine learning (ML) problem (Zeng et al., 2005; Shardlow, 2013). Indeed, before any gold-standard datasets were made available, the early approaches to the identification of complex words in a text included, on the one hand, readability measures determining complex words based on word familiarity (Dale and Chall, 1948) or on syllable count (Gunning, 1952; Mc Laughlin, 1969) and, on the other hand, simplification methods which plainly considered all words as complex and simplified every-

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 328–334
New Orleans, Louisiana, June 5, 2018.

thing (Devlin and Tait, 1998) or simplified words based on a threshold on word familiarity (Elhadad, 2006).

The SemEval-2016 shared task on complex word identification (described in detail in Paetzold and Specia, 2016a) was the first evaluation campaign which provided a gold-standard dataset as well as an extensive comparison of different machine learning approaches for the task at hand. The submitted systems included different types of classifiers such as SVMs, random forests, maximum entropy systems, ... which combined different types of features, ranging from linguistic information (on a lexical, morphological, semantic and syntactic level), over psycholinguistic measures to corpus-based information such as frequencies. The results on the shared task showed how ensemble methods (Paetzold and Specia, 2016b) outperformed any other ML technique and neural approaches in particular (Bingel et al., 2016). The task also showed however how a lack of annotation standards made it difficult for any ML-approach to model the rather inconsistent human assessment (Zampieri et al., 2017).

2.2 Deep Learning Architectures

The system we describe likewise inscribes itself in the ML-approach to CWI and draws inspiration from neural network literature in NLP. We adapt the architectures' initial purposes and apply it to the task at hand. Collobert et al. (2011) show how distributional information from words, called word embeddings, can be used in combination with a neural network architecture to largely replace hand-crafted features for learning NLP-related tasks such as POS-tagging and NER. The embeddings capture fine-grained information covering its linguistic behavior and the neural network model successfully teases out the relevant properties from that representation for the given task. Character embeddings (Zhang et al., 2015; Zhang and LeCun, 2015) take it one step further and also make it possible to encode and capture subword information in the modeling process.

3 Methods, Data, etc.

3.1 Data sources

The English datasets cover 3 informationally dense target domains for which to assess lexical complexity, namely news, Wikipedia and Wikipedia news. The Spanish dataset contains data taken from Spanish Wikipedia pages. Table 1 summarizes the number of training, development and test items for each dataset we used in the experiment. We combined training and development sets and used it as a single training set.

As a general domain corpus we use the COW-corpora (Schäfer, 2015; Schäfer and Bildhauer, 2012). The corpora are gathered online and cover a wide scope of topics. The English corpus contains well over 13 billion tokens, the Spanish one over 4 billion tokens.

We have at our disposal psychological measures for English from the MRC Psycholinguistic Database (Wilson, 1988). Measures include age of acquistion, imageability, concreteness, familiarity and meaningfulness and covers 150837 words. The overlap between the training dataset is however limited to approximately 1500 words.

Dataset	Train	Dev	Test
English News	14002	1764	2095
English Wikipedia	5551	694	870
English Wikinews	7746	870	1287
Spanish	13750	1622	2233

Table 1: CWI training, development and test sets

3.2 Feature operationalization

Psychological measures Psychological measures are used for the words found in the available dataset. Missing values were extrapolated based on findings that psychological measures correlate (inversely) to frequency. As such, less frequent words tend for instance to have a higher age of acquisition, and a lower imageability and concreteness rating. We therefore chose to respectively use third and first quartile values. In order to accommodate the neural network architecture all values have been normalized by dividing by the maximum value.

Frequency counts Frequency counts are calculated from the general corpus for all experiments. To avoid skewness we perform a rank transformation, with equal ranks being given the first encountered rank, and normalize again by dividing by the highest rank.

Word length Word length is also determined.

Word embeddings Word embeddings are pretrained using the COW-corpora and are used to

initialize several of our input layers in the neural network. We use the gensim implementation of word2vec to construct a 300 dimensional embedding space, based on a window-size of 5 including words that reach a minimum frequency threshold of 20.

Character embeddings Character embeddings are trained on the train and development set of all target words. Each character is replaced by a 16-dimensional encoding which has been randomly initialized. Each word consists of a concatenation of its character representations.

3.3 Architecture

Figure 1 shows the general architecture for the CWI-task. The model has been constructed with the Keras deep learning library (Chollet et al., 2015) with tensorflow-gpu as a backend. It includes the 5 sources of information we discussed in the previous section/ which are used as features to represent information at the word and the sentence level. At the word level, we include engineered features (psychological measures, corpus-counts and word length) and distributional information (word and character embeddings). At the sentence level we concatenate embeddings to capture topical information.

3.3.1 Input Layers

We include **engineered features** for the English dataset following the idea that they correlate with cognitive complexity. The features include psychological information, corpus-counts and word length. Corpus-counts measure familiarity and infrequent words are attributed a higher degree of complexity. Word length then has been shown to be related to processing difficulties and is relevant for instance to determine which words pose problems for persons with dyslexia.

Each target word is encoded by its **word embedding**, or in the case of word groups by their concatenated embeddings. The idea is that words with similar distributional patterns might have a comparable complexity. An LSTM layer with a dimensionality of 64 compacts the dimensionality of the representation.

Each target is also encoded as a sequence of its **character embeddings**. This input encoding is meant to capture morphological information as well as cues from letter sequences which might be perceived as difficult. The character embeddings are trained through 2 convolutional layers (4 filters, kernel size of 4, stride of 1) followed by max pooling (with a size of 2). An LSTM of size 64 is the final layer that directly encodes the character information.

The entire **sentence** is encoded as a concatenation of word embeddings and serves as a sort of topical approximation using contextual cues. An LSTM of 128 finalizes the information captured in this layer.

3.3.2 Dense Layers

All inputs are then concatenated and run through a shallow 3 layered fully connected network (each consisting of 32 nodes) with a moderate dropout rate of 0.3. A final dense layer predicts the output. 2 auxiliary loss functions are provided to ensure smooth training of the character and the topic model. We use binary cross-entropy as the loss function for the binary outcome task and mean squared error rate for the probabilistic one. We applied the architecture to the English datasets and, with the exception of the psychological measures, also to the Spanish one.

4 Results

Dataset	Result	Rank	Maximum-score
English News (Acc)	0.872	2	0.879
English Wikipedia (Acc)	0.782	5	0.812
English Wikinews (Acc)	0.815	6	0.843
Spanish (Acc)	0.777	2	0.784
English News (MAE)	0.054	2	0.051
English Wikipedia (MAE)	0.081	2	0.074
English Wikinews (MAE)	0.071	3	0.067
Spanish (MAE)	0.073	2	0.072

Table 2: Results, Rank and Maximum scores for the CWI identification task

The results in Table 2 show reasonably good performance for all tasks. Our architecture seems to work especially well for the regression task, but shows its aptitude for the classification task as well. The size of the training data seems to play a direct role in the system's ability for accurate predictions. This is in line with other deep learning literature. This does not hold for the Spanish set however, which might be due to a slight difference in apprehension during the data collection phase. The inclusion of corpus-counts and pre-trained embeddings from a general corpus, rather than a wikipedia corpus shows directly

Figure 1: Neural Network Architecture

Input	Precision Non-complex	Precision Complex	Recall Non-complex	Recall Complex
Character encoding (C)	0.876	0.757	0.839	0.809
Engineered features (E)	0.853	0.755	0.846	0.764
Word embeddings (W)	0.892	0.813	0.882	0.829
Sentences (S)	0.617	0	1	0
W + C	0.897	0.829	0.893	0.835
W + C + E	0.902	0.825	0.888	0.845

Table 3: Precision and Recall for different input layers

in the performance of the respective tasks. Using a wikipedia corpus will probably positively influence the results for those particular sets. Yet, the inclusion of general corpus-information proves to be a valid alternative in lack of specialized corpora. The inclusion of the engineered features does not seem to affect the obtained scores much.

Table 3 provides an overview of the relative contribution of each input layer to the final result for the English news dataset. The models were trained for 50 epochs. Considering each input layer separately, the word embeddings are the best estimator for the complexity task, followed closely by the character embeddings. Engineered features capture some information on the word's complexity, yet not as much as the embedding layers. Interestingly, sentence information does not outperform the baseline.

The combination of input layers shows the relative improvement that can be achieved by adding more information to the best performing input layer. The results indicate that combining information only marginally improves performance. They also confirm that the engineered features in combination with the embeddings do not contribute much to the final score.

This leads to the following conclusions for the current dataset. First, complexity is best determined by including focused information of the target word itself. The inclusion of contextual, topical information does not show any noticeable advantage. Looking at the combination of input layers, we can derive that the engineered features only add marginally different information from other input sources. This could be due to the limited number of words that are actually covered by the psychological dataset, but it also implies that the information from the corpus-counts is indirectly captured by the embeddings and from the word length by the character encodings. It is a case in point for replacing manual feature engineering by word and character embeddings. Based on these results we cannot conclude whether the word embeddings' better performance over the character embeddings is due to pre-training.

5 Conclusion

Reasonable performance can be achieved with a single architecture including information from different levels of linguistic description. Information derived from large scale corpora makes it possible to include them as a starting point on which to build a general architecture that learns the appropriate weights for the specific problem, in our case, the CWI-task. Embeddings at the word and the character level seem to contain sufficient information to model the problem well.

Future work will include an exploration to find optimal hyperparameter settings to optimize the identification task. We will likewise explore whether pre-training the character embeddings on a larger corpus will put its performance on par with the pre-trained word embeddings. The latter would pave the way for a model with less training parameters and would significantly reduce complexity.

References

Joachim Bingel, Natalie Schluter, and Héctor Martínez Alonso. 2016. CoastalCPH at SemEval-2016 Task 11: The importance of designing your Neural Networks right. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1028–1033, San Diego, California. Association for Computational Linguistics.

François Chollet et al. 2015. Keras. `https://keras.io`.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability. *Educational Research Bulletin*, 27(1):11–28.

Jan De Belder and Marie-Francine Moens. 2010. Text Simplification for Children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM.

Siobhan Devlin and John Tait. 1998. The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers. In John Nerbonne, editor, *Linguistic Databases*, number 77 in CSLI Lecture Notes, pages 161–173.

N. Elhadad. 2006. Comprehending technical texts: predicting and defining unfamiliar terms., Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms. *AMIA Annual Symposium Proceedings, AMIA Annual Symposium Proceedings*, 2006:239, 239–243.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

G. Harry Mc Laughlin. 1969. SMOG Grading—-a New Readability Formula. *Journal of Reading*, 12(8):639–646.

Gustavo Paetzold and Lucia Specia. 2016a. 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. 2016b. 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.

Gustavo Henrique Paetzold and Lucia Specia. 2016c. Understanding the Lexical Simplification Needs of Non-Native Speakers of English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727, Osaka, Japan.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or Help?: Text Simplification Strategies for People with Dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 15:1–15:10, New York, NY, USA. ACM.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster. UCREL, IDS.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool ChainNo Title. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2014. Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline. In *LREC 2014*, pages 1583–1580. European Language Resources Association (ELRA).

Sanja Štajner, Victoria Yaneva, Ruslan Mitkov, and Simone Paolo Ponzetto. 2017. Effects of Lexical Properties on Viewing Time per Word in Autistic and Neurotypical Readers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 271–281, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Wilson. 1988. MRC Psycholinguistic Database: Machine Usable Dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

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 Recent Advances in Natural Language Processing*, pages 813–822, Varna, Bulgaria. Incoma Ltd. Shoumen, Bulgaria.

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.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A Text Corpora-Based Estimation of the Familiarity of Health Terminology. In *Biological and Medical Data Analysis*, Lecture Notes in Computer Science, pages 184–192. Springer, Berlin, Heidelberg.

Xiang Zhang and Yann LeCun. 2015. Text Understanding from Scratch. *arXiv:1502.01710 [cs]*. ArXiv: 1502.01710.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. pages 1–9.