Complex Word Identification: Convolutional Neural Network vs. Feature Engineering

Segun Taofeek Aroyehun, Jason Angel, Daniel Alejandro Pérez Alvarez, Alexander Gelbukh

aroyehun.segun@gmail.com, ajason08@gmail.com, daperezalvarez@gmail.com,www.gelbukh.com

CIC, Instituto Politécnico Nacional Mexico City, Mexico

Problem

- Complex words inhibit the reading comprehension of different target audience such as non-native speakers, and native speakers with cognitive impairments
- Complex Word Identification (CWI) is the ability to identify word(s) as complex or not in a given context
- CWI is an important step in text simplification
- The organizers of the 2018 CWI shared task [1] provided participants with multilingual human-annotated datasets [2, 3] for the identification of complex words
- We developed classifiers for CWI using two approaches: feature engineering and CNN

Model 1: Feature engineering

Features

- **Morphological Features:** frequency count of target text in Wikipedia and Simple Wikipedia, number of characters, vowels and syllables
- **Syntactic and Lexical Features:** part-of-speech (POS) tag, and number of senses, lemmas, hyponyms, hyperonyms
- **Psycholinguistic and Entity Features:** familiarity, age of acquisition, concreteness, and imagery plus entity tags
- **Word Embedding Distances as Features:** cosine distance between the average of the vector representation of the words (pre-trained word2vec) in the sentence and the target text

Classical Machine Learning Models

- Tree learner performed better than other classical machine learning models
- The best obtained result was given by the tree ensembles with 600 models

References

[1] Yimam, Seid Muhie and Biemann, Chris and Malmasi, Shervin and Paetzold, Gustavo and Specia, Lucia and Štajner, Sanja and Tack, Anaïs and Zampieri, Marcos: *A Report on the Complex Word Identification Shared Task 2018*, Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications, New Orleans (2018)

[2] Yimam, Seid Muhie and Štajner, Sanja and Riedl, Martin and Biemann, Chris: *CWIG3G2-Complex Word Identification Task across Three Text Genres and Two User Groups*, Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (2017)

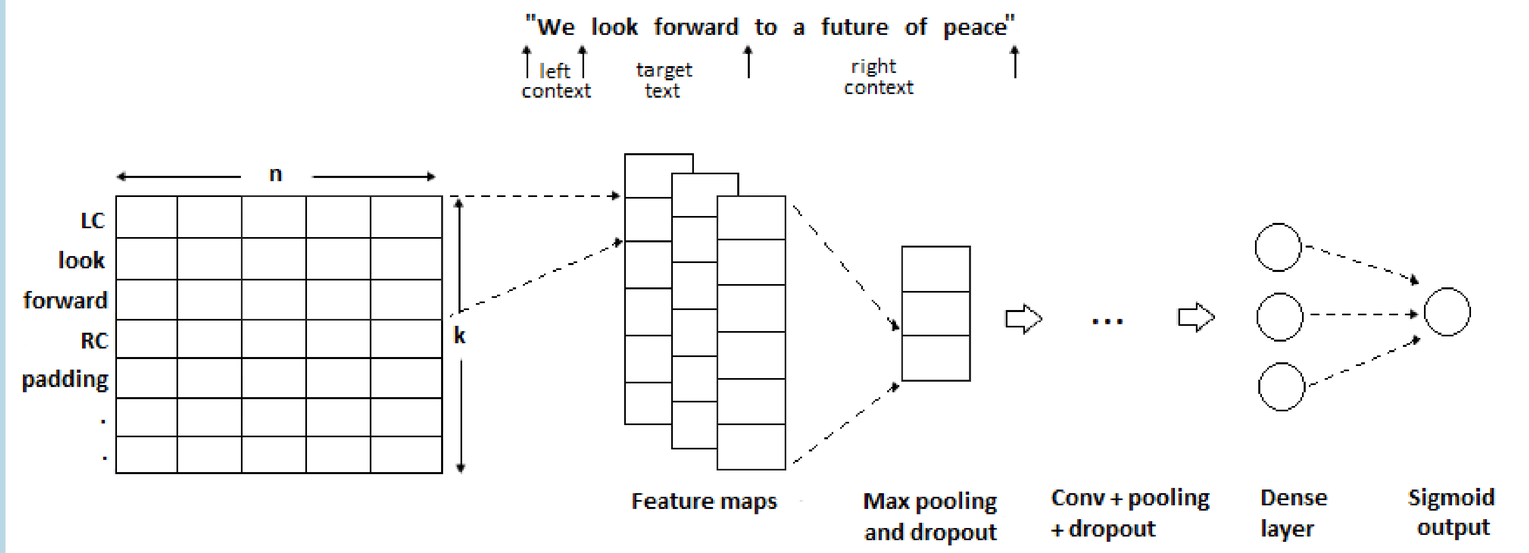
[3] Yimam, Seid Muhie and Štajner, Sanja and Riedl, Martin and Biemann, Chris: *Multilingual and Cross-Lingual Complex Word Identification*. Proceedings of RANLP (2017)

Acknowledgements

The support of the Mexican government via CONACYT (SNI) and the Instituto Politécnico Nacional grant SIP-20181792 is gratefully acknowledged.

Model 2: CNN

- Word embedding representation (word2vec for English, fastText for Spanish)
- Context representation as average of word vectors
- CNN using the vector representation of the target text and context as input
- We trained our model with dropout (0.25) and earlystopping for 100 epochs



Results

- The CNN and Tree ensemble showed comparable performance on the English test set (Table 1)
- Both models are within 0.01 of the system with the best macro-F1
- The CNN model ranked third on the spanish test set (Table 2)
- Table 3 shows the sensitivity of both models on the English test set to the number of characters

Models	News			Wikinews			Wikipedia		
	Macro-F1	Accuracy	Rank	Macro-F1	Accuracy	Rank	Macro-F1	Accuracy	Rank
NLP-CIC-TreeE	0.851	0.859	9	**0.831**	**0.837**	3	0.772	**0.774**	11
NLP-CIC-CNN	**0.855**	**0.863**	8	0.824	0.828	7	0.772	0.772	12

Table 1: Performance on the English Test set

Model	Macro-Recall	Macro-Precision	Macro-F1	Accuracy	Rank
NLP-CIC-CNN	0.765	0.772	0.767	0.772	3

Table 2: CNN Performance Scores on the Spanish test set

Source	NLP-CIC-TreeE Model		NLP-CIC-CNN Model	
	Correct	Wrong	Correct	Wrong
Wikinews	0.94 ± 0.53	1.10 ± 0.65	0.94 ± 0.51	1.12 ± 0.72
News	0.97 ± 0.55	1.21 ± 0.75	0.97 ± 0.55	1.17 ± 0.75
Wikipedia	1.05 ± 0.65	1.04 ± 0.68	1.04 ± 0.66	1.08 ± 0.65

Table 3: Model Performance Sensitivity to character count on the English Test set

Conclusions

- The Tree ensemble and CNN showed comparable performance
- For the English track, our best model placed fifth on News, second on Wikinews, and seventh on Wikipedia
- The CNN model can be successfully applied to another language given the availability of pre-trained embedding
- The CNN model ranked third overall on the Spanish test set
- Our models tend to fail on longer target texts
- The impact of domain-specific features will be evaluated in the future