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, hypernyms, hyponyms
- **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

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

<table>
<thead>
<tr>
<th>Models</th>
<th>English Test Set</th>
<th>Spanish Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP-CIC-TreeE</td>
<td>0.851</td>
<td>0.853</td>
</tr>
<tr>
<td>NLP-CIC-CNN</td>
<td>0.855</td>
<td>0.863</td>
</tr>
</tbody>
</table>

Table 1: Performance on the English Test set

<table>
<thead>
<tr>
<th>Model</th>
<th>Macro-Recall</th>
<th>Macro-Precision</th>
<th>Macro-F1</th>
<th>Accuracy</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLP-CIC-CNN</td>
<td>0.765</td>
<td>0.772</td>
<td>0.767</td>
<td>0.772</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2: CNN Performance Scores on the Spanish test set

<table>
<thead>
<tr>
<th>Source</th>
<th>NLP-CIC-TreeE Model Correct</th>
<th>NLP-CIC-TreeE Model Wrong</th>
<th>NLP-CIC-CNN Model Correct</th>
<th>NLP-CIC-CNN Model Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikinews</td>
<td>0.94 ± 0.53</td>
<td>1.10 ± 0.65</td>
<td>0.94 ± 0.51</td>
<td>1.12 ± 0.72</td>
</tr>
<tr>
<td>News</td>
<td>0.97 ± 0.55</td>
<td>1.21 ± 0.75</td>
<td>0.97 ± 0.55</td>
<td>1.17 ± 0.75</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1.05 ± 0.65</td>
<td>1.04 ± 0.68</td>
<td>1.04 ± 0.66</td>
<td>1.08 ± 0.65</td>
</tr>
</tbody>
</table>

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

References

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