(1) Motivation

Example: **Hundreds of protesters** have been detained in both cities.
- hundreds ~ more than 100
- protesters ~ demonstrators
- detained ~ arrested
- cities ~ ?

CWI is an important task in its own right:
1. facilitates more targeted text adaptation
2. helps avoid unnecessary & educationally harmful “oversimplification”
3. alleviates data sparsity: definition can be provided if no simpler alternative available

(2) CWI Shared Task 2018

Data:
- 3 data sources: News (NEWS), WikiNews (WNI) and Wikipedia (WIKI)
- Content words and phrases annotated via MTurk by 10 native and 10 non-native speakers (metadata not used or released)
- 2 settings: bin if at least one annotation as CW(prob) = proportion of annotators

Annotation:

<table>
<thead>
<tr>
<th>Data</th>
<th>∅(bin)</th>
<th>1(bin)</th>
<th>0.05(prob)</th>
<th>1.0(prob)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEWS</td>
<td>60.41</td>
<td>39.59</td>
<td>13.52</td>
<td>0.39</td>
</tr>
<tr>
<td>NEWS</td>
<td>60.54</td>
<td>39.46</td>
<td>13.83</td>
<td>0.28</td>
</tr>
<tr>
<td>NEWS</td>
<td>61.72</td>
<td>38.28</td>
<td>12.70</td>
<td>0.29</td>
</tr>
<tr>
<td>WNI</td>
<td>58.48</td>
<td>41.52</td>
<td>16.25</td>
<td>0.17</td>
</tr>
<tr>
<td>WNI</td>
<td>59.43</td>
<td>40.57</td>
<td>14.25</td>
<td>0.11</td>
</tr>
<tr>
<td>WNI</td>
<td>57.58</td>
<td>42.42</td>
<td>16.71</td>
<td>0.16</td>
</tr>
<tr>
<td>WIKI</td>
<td>55.07</td>
<td>44.93</td>
<td>16.66</td>
<td>0.52</td>
</tr>
<tr>
<td>WIKI</td>
<td>51.15</td>
<td>48.85</td>
<td>19.31</td>
<td>0.14</td>
</tr>
<tr>
<td>WIKI</td>
<td>49.54</td>
<td>50.46</td>
<td>18.62</td>
<td>0.23</td>
</tr>
</tbody>
</table>

(3) Challenges

1. Context-specific annotation: up to 10% words receive different annotations; e.g., tragedy from 0.90 to 1.00 – interaction of surrounding context, multiple senses and sequence labelling effect
2. Sequence labelling effect: Beethoven’s Symphony No.7, Bruckner’s Symphony No.6 and Mendelssohn’s Symphony No.4 comprise a nearly complete list of symphonies in this key in the Romantic era.
3. Phrase annotation:
   - future_in0.05 ∪ generations0.25 = future generations
   - traditional0.2 ∪ connection0.0 ∪ country0.05 ≠ traditional connection to that country0.0
4. Proper nouns: 0.0 – 0.45 for Eurozone, 0.0 – 0.05 for Barack, 0.05 – 0.3 for Brexit, and 0.0 – 0.1 for Copenhagen, Estonia, Hungary, Warsaw, etc.

(4) CAMB System Overview

- Preliminary experiments confirm that ensemble-based approaches work best
- Method for bin setting:
  - WIKIPEDIA & NEWS – AdaBoost with 5,000 estimators
  - WikiNEWS – ensemble voting classifier using AdaBoost and Random Forest
- Method for prob setting: Linear Regression; round the classifier’s predictions to the nearest value on [0.0, ..., 1.0] with the step of 0.05

(5) Experimental Results

Features Overview:
1. Word N-grams and PoS: words, character bigrams and PoS tags
2. Lexical Features: word length, number of syllables, number of senses, hyper- and hyponyms from the WordNet
3. Dependency Parse Relations: number of dependency relations for the target word
4. Lexicon-Based Features: presence/absence in the SubIMDB, the Simple Wikipedia and Ogden’s Basic English list, CEPR level from the Cambridge Advanced Learners Dictionary
5. Word Frequency in the Google N-grams
6. Psycholinguistic Features from the MCR Psycholinguistic Database: word familiarity, imageability, concreteness, age of acquisition, etc.

Feature selection:
- bin: all features for NEWS & WNI; all but MCR features for WIKI
- prob: all features for NEWS; all but MCR features for WNI & WIKI

(6) Analysis

- Per-Genre Performance: Unique words

<table>
<thead>
<tr>
<th></th>
<th>NEWS</th>
<th>WNI</th>
<th>WIKI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>13,401</td>
<td>7,559</td>
<td>5,439</td>
</tr>
<tr>
<td>Unique</td>
<td>3,376</td>
<td>3,334</td>
<td>1,517</td>
</tr>
<tr>
<td>%</td>
<td>25.08</td>
<td>44.16</td>
<td>58.44</td>
</tr>
</tbody>
</table>

- Classifiers in both settings perform best on NEWS: NEWS contains lowest number of complex words & lowest number of unique words – less challenging
- Wiki – more challenging for humans (highest CW %) and machines (lowest results) + highest number of unique words

- Phrase Classification

<table>
<thead>
<tr>
<th>Data</th>
<th>Acc</th>
<th>P</th>
<th>R</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CW pres.</td>
<td>0.8094</td>
<td>0.8015</td>
<td>0.9977</td>
<td>0.8889</td>
</tr>
<tr>
<td>N-gram</td>
<td>0.8004</td>
<td>0.8004</td>
<td>1.0000</td>
<td>0.8961</td>
</tr>
<tr>
<td>Greedy*</td>
<td>0.8004</td>
<td>0.8004</td>
<td>1.0000</td>
<td>0.8961</td>
</tr>
</tbody>
</table>

(7) Conclusions

Our systems scored first on all 3 text genres in the bin classification track, and on 2 out of 3 genres in the prob track. Further analysis identifies future directions for this research.

1. contextualisation of CWI
2. better phrase complexity prediction
3. personalisation of CWI with level of education, L1 and level of language competence

Contact Information

Email: Sian.Gooding@cl.cam.ac.uk
Web: www.cl.cam.ac.uk/~ek388/