Lexical Complexity Prediction of English words using the Complex Corpus

By Abhinandan Desai

Advisor: Dr. Marcos Zampieri

Faculty Advisor: Dr. Christopher Homan
<table>
<thead>
<tr>
<th>CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PROJECT OVERVIEW</td>
</tr>
<tr>
<td>MILESTONE 1 OVERVIEW</td>
</tr>
<tr>
<td>MILESTONE 2 GOALS</td>
</tr>
<tr>
<td>EXPLORATORY ANALYSIS</td>
</tr>
<tr>
<td>EXPERIMENTAL RESULTS</td>
</tr>
<tr>
<td>EVALUATION</td>
</tr>
<tr>
<td>MILESTONE 3 GOALS</td>
</tr>
<tr>
<td>REFERENCES</td>
</tr>
</tbody>
</table>
PROJECT OVERVIEW

Enhancing Lexical Complexity Prediction (LCP)

CompLex corpus – 5-point Likert scale annotation scheme

Simplify texts for various target audiences

Performance analysis of the systems
Dataset Description

Data Preprocessing and Preparation

Feature Calculation:
- Length of token
- Number of Syllables in the token
- Frequency of the token

Figure 1: Single word Dataset

Figure 2: Single word features
MILESTONE 2 GOALS

Explore and analyze the calculated features

Work on ways to analyze the results

Develop ML architecture to analyze features

Research into word context models
EXPLORATORY ANALYSIS

- The frequency of the tokens were generalized for English language
- Reference Corpora:
  - Brown Corpus[1] – More than 1 million words sampled from 15 text categories
  - Reuters Corpus[2] - 800,000 news stories in English language

<table>
<thead>
<tr>
<th>token</th>
<th>wordLength</th>
<th>syllableCount</th>
<th>frequency</th>
<th>complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>river</td>
<td>5</td>
<td>2</td>
<td>274</td>
<td>0.000000</td>
</tr>
<tr>
<td>brothers</td>
<td>8</td>
<td>2</td>
<td>151</td>
<td>0.000000</td>
</tr>
<tr>
<td>sons</td>
<td>4</td>
<td>1</td>
<td>41</td>
<td>0.160714</td>
</tr>
<tr>
<td>days</td>
<td>4</td>
<td>1</td>
<td>876</td>
<td>0.000000</td>
</tr>
<tr>
<td>lion</td>
<td>4</td>
<td>2</td>
<td>20</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

Figure 3: Single word features with frequency updated
EXPLORATORY ANALYSIS

**Figure 4:** Frequency plot for single word

**Figure 5:** Frequency plot for Multi-word expressions
EXPLORATORY ANALYSIS (PROBLEM)

Figure 6: Word Length plot for single word

Figure 7: Word length plot for Multi-word expressions

Figure 8: Syllable count plot for single word

Figure 9: Syllable count plot for Multi-word expressions
EXPLORATORY ANALYSIS

**Figure 10:** Word length boxplot for single word

**Figure 11:** Word length boxplot for Multi-word expressions
EXPLORATORY ANALYSIS

Figure 12: Syllable count boxplot for single word

Figure 13: Syllable count boxplot for Multi-word expressions
The baseline system is a **Linear Regression Model**

- **Data Distribution:**
  - Training Data – 70 percent
  - Test Data – 30 percent

- The model predicts complexity value

**Figure 14**: Linear regression results for single word

**Figure 15**: Linear regression results for Multi-word expressions
EVALUATION CRITERIA

- **Mean Absolute Error (MAE)** – preferred when variable data is continuous
  
  \[ MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_{pred}| \]

- **Mean Squared Error (MSE)** – preferred when data contains outliers and unexpected values
  
  \[ MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - y_{pred})^2 \]

- **Root Mean Squared Error (RMSE)** – assigns higher weight to larger errors
  
  \[ RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - y_{pred})^2} \]
EVALUATION RESULTS

- Linear Regression performed better when the token was a single word.
- Both datasets have less outliers as MSE values are low.
- High RMSE values indicate larger errors are present.

**Figure 16**: Evaluation results for single word

<table>
<thead>
<tr>
<th>Mean Absolute Error</th>
<th>Mean Squared Error</th>
<th>Root Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.102</td>
<td>0.017</td>
<td>0.13</td>
</tr>
</tbody>
</table>

**Figure 17**: Evaluation results for Multi-word expressions

<table>
<thead>
<tr>
<th>Mean Absolute Error</th>
<th>Mean Squared Error</th>
<th>Root Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.114</td>
<td>0.019</td>
<td>0.139</td>
</tr>
</tbody>
</table>
MILESTONE 3 GOALS

- Regression models for comparative analysis
- Kendall Tau Distance as an evaluation criteria
- Final Report

THANK YOU!