Measuring Feature Diversity in Native Language Identification

Shervin Malmasi
Macquarie University
Australia

Aoife Cahill
Educational Testing Service
USA
ML for NLI

• Predicting the native language of a writer based on a piece of English writing
• Typically solved using supervised-ML: multi-class classification
• Previous Work has investigated the predictive power of individual feature classes
• No systematic analysis of feature interaction
Beyond NLI System Performance

• Context: language teaching and learning
• Goal: identify L1-specific usage patterns and errors
• Improve teaching methods, instructions and learner feedback
• Previous work shows that the features capture different pieces of information
• How diverse are the features? How can we measure the diversity?
# Feature Types for NLI

<table>
<thead>
<tr>
<th><strong>Lexical</strong></th>
<th><strong>Syntactic</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>character n-grams</td>
<td>POS n-grams</td>
</tr>
<tr>
<td>word n-grams</td>
<td>syntactic dependencies</td>
</tr>
<tr>
<td>lemma n-grams</td>
<td>TSG fragments</td>
</tr>
<tr>
<td>function words</td>
<td>CFG rules</td>
</tr>
<tr>
<td></td>
<td>Adaptor grammars</td>
</tr>
</tbody>
</table>
Data

• ETS Corpus of Non-Native English Writing (TOEFL 11)
• 11 L1s: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu, Turkish
• 1100 essays per L1, 900 train, 100 dev/test
• 8 prompts
• Train on train+dev, Evaluate on test
Accuracy of Individual Features
Measuring Feature Diversity

• Measure agreement between each pair of features for predicting labels on the same dataset

• Idea: the higher the agreement, the lower the diversity of those two features

• Yule’s Q-coefficient statistic
Yule’s Q-coefficient

- Correlation coefficient for binary measurements
- Range from -1 to +1

\[ Q_{j,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} \]
Q-coefficients (171 pairs)
Q-coefficient Matrix
Words and Dependencies

• Naively not thought to be strongly related
• Liu (2008) reports 51% of deps are adjacent
• How does this relate to k-skip word bigrams?
Q-coefficient Matrix
L1 and Word Usage

• Hypothesis: learners tend to use words similar in form and meaning to words in their L1
• Test: Extract English words from Etymological WordNet
  – Germanic roots
  – Latin roots
• Train 2 classifiers with just word unigrams
  – 2 SVMs each trained on different features
L1 and Word Usage Results

![Graph showing L1 and Word Usage Results for various languages, comparing Old English Word Features and Latin Word Features.](image)
Q-coefficient Matrix
Extending CFG Rules

• Parent Annotations (Johnson, 1998)

Production Rules Extracted from Tree:

ROOT → S
NP → DT JJ JJ NN
VP → VBD PP
S → NP VP .
PP → IN NP
NP → DT JJ NN
VBD
IN
NP
over
the lazy dog
The quick brown fox jumped
Parent-Annotated CFG Rules

\[
\begin{align*}
\text{ROOT} & \rightarrow S^{^<\text{ROOT}>} \\
S^{^<\text{ROOT}>} & \rightarrow NP^{^<S>} \; VP^{^<S>} \; . \\
NP^{^<S>} & \rightarrow DT \; JJ \; JJ \; NN \\
VP^{^<S>} & \rightarrow VBD \; PP^{^<VP>} \\
PP^{^<VP>} & \rightarrow IN \; NP^{^<PP>} \\
NP^{^<PP>} & \rightarrow DT \; JJ \; NN
\end{align*}
\]

Building an NLI system with these features yields accuracy of 55.6%, a +1.3% increase over the standard CFG rules feature.
Conclusions

• Q-coefficient provides a method for measuring feature diversity for high-dimensional feature spaces

• Experiments with NLI on TOEFL data show interesting feature correlations

• Analysis of feature diversity can help suggest new features