Combining Shallow and Linguistically Motivated Features in Native Language Identification

Serhiy Bykh, Sowmya Vajjala, Julia Krivanek and Detmar Meurers
Department of Linguistics, University of Tübingen

In a Nutshell

- We explored a wide range of features:
  - from surface (e.g., n-grams)
  - to deeper linguistic features (e.g., dependency)
- We created ensemble classifiers by combining multiple single-feature classifiers, significantly increasing performance.
- Our best accuracy of 83.5% is the second best score in the overall ranking of the NLI Shared Task (Tetreault et al., 2013).
  - Closed task: 82.2% (rank 5, difference to best result 83.6% not statistically significant)
  - Open-2 task: 83.5% (rank 1)
  - Open-1 task: 38.5% (rank 2)

Background

- Early work on NLI has explored different kinds of features ranging from word n-grams to spelling and grammar errors (e.g., Jarvis et al., 2004; Koppel et al., 2005)
- Wong & Dras (2009) used features based on Contrastive Analysis.
- More recently, complex syntactic constructs were used as features (e.g., Wong & Dras, 2011; Swanson & Charniak, 2012)
- Brooke & Hirst (2011) studied the effect of training data size on classifier performance.
- Tetreault et al. (2012) used ensemble models that combine multiple classifier performance.
- More recently, complex syntactic constructs were used as features.
- Bykh & Meurers (2012) explored a data driven approach using feature groups by building a meta-classifier of base classifiers.
- Tetreault et al. (2012) used ensemble models that combine multiple classifier performance.
- More recently, complex syntactic constructs were used as features.

Corpora used

TOEFL11 (Blanchard et al., 2013)
- Main corpus of the shared task
- 1100 essays of English learners with 11 L1 backgrounds.

NON-TOEFL11
- 5843 essays for 11 L1s for the open-1 and open-2 tasks
- unevenly distributed across 11 L1s, created from 5 corpora:
  - ICLE corpus (Granger et al., 2009)
  - FCE corpus (Yannakoudakis et al., 2011)
  - BALC Arabic Learner Corpus (Randall & Groom, 2009)
  - ICNALE corpus (Ishikawa, 2011)
  - TÜTEL-NLI: Tübingen Telugu NLI Corpus

Features

Recurring n-gram features

1. rc. word ng. recurring-word-based n-grams
2. rc. OCP-pos ng. recurring n-grams, where open class words are replaced by POS tags
3. rc. word dep. rec. word-based dependencies (MATE): a head and all its immediate dependents, ordered as in the sentence
4. rc. func. dep. rec. function-based dependencies: each dependent is replaced by its grammatical function

Complexity Features

5. complexity text complexity features of Vajjala & Meurers (2012): lexical richness, syntactic complexity...
6. morphological and POS features from CELEX

Sublexical Morphological Features

6. stem-suffix, bin. presence/absence of stem-suffix.
7. stem-suffix, cnt. number of stem-suffix occurrences.
8. suffix, bin. presence/absence of valid English suffixes.
9. suffix, cnt. number of suffix occurrences.

Constituency Parser-based Features

10. type dep. lm. lemma-type Stanford dependencies
11. type dep. POS POS-based Stanford dependencies
12. local trees all syntactic trees of depth one

Ratio Features

13. dep. num. number of dependents (MATE) realized by a verb
14. dep. var. number of possible dependent POS combinations for a verb
15. dep. POS POS-based dependent frequency for a verb
16. lem. realiz. lemma counts of a specific POS normalized by the total count of this POS

Discussion

- Best single feature group: surface-based recurring n-grams
- Ensemble models combining a range of linguistically motivated features clearly outperform individual feature models.
  - Even individually weak features significantly contribute.

Future Work

- Qualitatively analyze feature types in depth and study the correlations between them.
- Explore more linguistic features like syntactic alternations as proposed in Krivanek (2012)

Experimental setup & Results

- We submitted five system results for each of the three tasks.
- The ensembles are meta-classifiers created based on the probability distributions of the base classifiers.
- All systems consisted of classifier ensembles, except system 2.

<table>
<thead>
<tr>
<th>Feature type</th>
<th>systems</th>
<th>Single feature results on T11 dev set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>closed</td>
<td>open1</td>
</tr>
<tr>
<td>1. rc. word ng.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2. rc. OCP-pos ng.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3. rc. word dep.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>4. rc. func. dep.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>5. complexity</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

The starred Open task results finished computing after submission.

Task | Overall system results
Closed| 82.2 | 79.6 | 81.0 | 81.5 | 74.7
Closed-dev | 85.4 | 81.3 | 83.5 | 84.9 | 76.3
Closed-dev | 82.4 | 78.9 | 80.7 | 81.7 | 74.1
Open1-dev | 36.4 | 38.5 | 33.2 | 57.8 | 21.2
Open1-dev* | 37.0 | 38.5 | 35.4 | 57.8 | 29.9
Open2-dev | 83.5 | 81.0 | 79.3 | 82.5 | 64.8
Open2-dev* | 84.5 | 81.0 | 83.3 | 82.9 | 79.8