Using Other Learner Corpora in the 2013 NLI Shared Task

1. Introduction
- The 2013 Native Language Identification Shared Task (Tetreault et al. 2013)
- Participated in all three tasks: one closed-training, two open-training
- The new TOEFL-11 learner essay corpus (Blanchard et al. 2013)
- Well controlled, but limited in scope?
- Our focus is on building robust models
- Use of cross-linguistic evaluation (Brooke and Hirst 2012; Bykh and Meurers 2012)

2. Base Model
- Training Corpus
- Full Parse
- Feature Extraction
- Feature/Data Selection
- Elements in blue used only in open-training task
- Statistical Classifier
- Bias Adaption

3. Closed Task
- Standard features (Brooke and Hirst 2012)
- Character trigrams
- POS trigrams
- Context-free grammar productions
- Only dependencies improved model
- Feature selection
- Feature frequency cutoff
- No feature selection preferred
- New features
- Partial abstractions
- Dependency chains
- Productions with unspecified elements
- TSG fragments (Swanson and Charniak 2013)
- MRC psycholinguistic lexicon (Coltheart, 1980)
- All new features fail to improve on best from Table 2

4. External Corpora for Open-Training Tasks
- Lang-8 Corpus (Brooke and Hirst 2012)
- Noisy but good coverage web corpus
- ICLC (Granger et al., 2009)
- FCE (Yannakoudakis et al., 2011)
- Short answers (letters, short stories)
- ICCE (Tono et al., 2012)
- Grade school essays
- ICNALE (Ishikawa, 2011)
- Asian controlled for genre

5. Open-Training Task 2
- Failed attempt: Meta-classifier
- Worse than TOEFL-11 alone
- Best score 78.5%
- Main approach: Combine data
- Method 1: Bias adaption (BA)
  - Equalize output class ratios
  - Result: Lang-8 definitely helps

6. Open-Training Task 1
- Approach to blind task: testing on other corpora
- Effect of bias adaption is key
- Corpus selection inconsistent
- Again, all corpora useful for training except ICCE
- Efficiency profits?
- Post-hoc analysis in TOEFL-11
  - confirms that all but the ICCE are useful and domain adaption helps
  - All three Indian corpora can distinguish Hindi/Telugu somewhat, but only tweets and news are useful additions for NLI in TOEFL-11

References and Acknowledgements

Table 1: Feature testing for closed-training task, post-only investigated features; best result in bold

Table 2: Feature frequency testing for closed-training task; best result in bold

Table 3: Number of tokens (in thousands) in Indian corpora, by L1

Table 4: Number of tokens (in thousands) in learner corpora, by L1

Table 5: Number of tokens (in thousands) in Indian corpora

Table 6: Corpus testing for open-training task 1; best result in bold

Table 7: Training Set selected for open-training task 1; best result in bold

Table 8: Corpus testing for open-training task 1; best result in bold

Table 9: 11-language testing on TOEFL-11 area for open training task 1; best result in bold, submitted run in Table 10

Table 10: Indian corpus testing for open-training task 1; best result in bold

© Copyright 2013 Julian Brooke and Graeme Hirst