Julian Brooke and Graeme Hirst

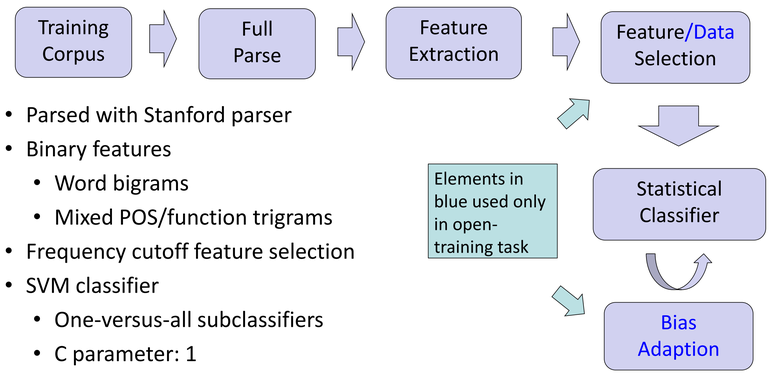
jbrooke, gh@cs.toronto.edu Department of Computer Science, University of Toronto

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-corpus evaluation (Brooke and Hirst 2012; Bykh and Meurers 2012)

2. Base Model



- Parsed with Stanford parser
- Binary features
  - Word bigrams
  - Mixed POS/function trigrams
- Frequency cutoff feature selection
- SVM classifier
  - One-versus-all subclassifiers
  - C parameter: 1

3. Closed Task

- Standard features (Brooke and Hirst 2012)
  - Character trigrams
  - POS trigrams
  - Context-free grammar productions
  - Dependencies
- Only dependencies improved model

Table 1: Feature testing for closed-training task, previously investigated features; best result is in bold.

Feature Set	Accuracy (%)
Word+mixed	76.8
Word+mixed+characters	72.0
Word+mixed+POS	76.6
Word+mixed+productions	77.9
Word+mixed+dependencies	**78.9**
Word+mixed+dep+prod	78.4

- Feature selection
  - Frequency cutoff
- No feature selection preferred

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

Cutoff	Accuracy (%)
At least 5 occurrences	78.9
At least 3 occurrences	79.5
At least 2 occurrences	79.7
All features	**80.2**

- 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

Table 3: Feature testing for closed-training task, new features; best result is in bold.

Feature Set	Accuracy (%)
Best	**80.2**
Best+partial abstraction	79.7
Best+dependency chains	78.6
Best+wild card productions	78.8
Best+TSG fragments	78.1
Best+MRC lexicon	54.2

4. External Corpora for Open-Training Tasks

- Lang-8 Corpus (Brooke and Hirst 2012)
  - Noisy but good coverage web corpus
- ICLE (Granger et al., 2009)
- FCE (Yannakoudakis et al., 2011)
  - Short answers (letters, short stories)
- ICCI (Tono et al., 2012)
  - Grade school essays
- ICNALE (Ishikawa, 2011)
  - Asian college essays
  - Strictly controlled for genre

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

L1	Corpus				
	Lang-8 (new)	ICLE	FCE	ICCI	ICNALE
Japanese	11694k	227k	33k	232k	199k
Chinese	7044k	552k	30k	243k	366k
Korean	5174k	0k	37k	0k	151k
French	536k	256k	61k	0k	0k
Spanish	861k	225k	83k	49k	0k
Italian	450k	251k	31k	0k	0k
German	331k	258k	29k	91k	0k
Turkish	51k	222k	22k	0k	0k
Arabic	218k	0k	0k	0k	0k
Hindi	11k	0k	0k	0k	0k
Telugu	2k	0k	0k	0k	0k

Table 5: Number of tokens (in thousands) in Indian corpora, by expected L1.

L1	Indian Corpus		
	News	Twitter	Blog
Hindi	996k	146k	2089k
Telugu	998k	133k	76k

- Few Hindi and Telugu texts
- Solution: Indian corpora
  - News from Hindi/Telugu areas
  - Tweets geolocated in these areas
  - Translated ICWSM blog posts (Burton et al. 2009)

5. Open-Training Task 2

- Failed attempt: Metaclassifier
  - 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

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

Training Set	Accuracy (%)	
	no BA	with BA
TOEFL-11 only	79.7	79.2
+Lang-8	79.5	**80.5**
+ICLE	80.2	80.2
+FCE	79.6	79.3
+ICCI	77.3	76.7
+ICANLE	79.7	79.3
+Lang-8+ICLE	80.4	80.4
+all but ICCI	80.0	80.4

- Method 2: Training data selection
  - Rank training data on the basis of test data language model
  - Remove fraction *r* of data
- Result: all data but ICCI is useful when both methods applied

Table 7: Training set selection testing for open-training task 2; best result is in bold, best submitted run is in italics.

Training Set	Accuracy (%)	
	no BA	with BA
TOEFL-11 only	79.7	79.2
+Lang-8	79.5	80.5
+Lang-8 $r = 0.1$	81.4	81.6
+Lang-8 $r = 0.2$	80.6	81.5
+Lang-8 $r = 0.3$	81.0	80.6
+all but ICCI	80.0	80.4
+all but ICCI $r = 0.1$	81.5	**82.5**
+all but ICCI $r = 0.2$	81.0	*81.6*
+all but ICCI $r = 0.3$	80.9	81.3

6. Open-Training Task 1

Table 8: ICLE testing for Open-training task 1; best result is in bold.

Training Set	Accuracy (%)	
	no BA	with BA
Lang-8	47.0	57.1
Lang-8+FCE	47.9	58.2
Lang-8+ICCI	46.4	54.8
Lang-8+ICNALE	46.9	57.5
Lang-8+ICNALE+FCE	47.7	**58.8**
Lang-8+ICNALE+FCE $r = 0.1$	46.6	58.2

Table 9: ICNALE testing for open-training task 1; best result is in bold.

Training Set	Accuracy	
	no BA	with BA
Lang-8	37.2	59.6
Lang-8+FCE	37.9	61.3
Lang-8+ICCI	35.7	61.4
Lang-8+ICLE	37.3	61.4
Lang-8+ICLE+FCE	37.6	61.7
Lang-8+ICLE+FCE $r = 0.1$	37.7	**61.9**

- Approach to blind task: testing on other corpora
- Effect of bias adaption is key
- Corpus selection inconsistent
- Again, all corpora useful for training except ICCI
  - Proficiency effects?
- Post-hoc analysis in TOEFL-11
  - confirms that all but the ICCI 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

Table 11: 11-language testing on TOEFL-11 sets for open-training task 1; best result is in bold, best submitted run is in italics.

Training Set	Accuracy (%)			
	TOEFL-11 test		TOEFL-11 training	
	no BA	with BA	no BA	with BA
Lang-8	39.5	53.2	37.2	48.2
Lang-8+ICCI	36.9	51.0	34.9	46.3
Lang-8+FCE+ICLE+ICNALE	44.5	55.8	44.9	53.1
Lang-8+FCE+ICLE+ICNALE+Indian news	45.2	*56.5*	45.5	54.9
Lang-8+FCE+ICLE+ICNALE+Indian tweets	44.9	56.4	45.1	53.4
Lang-8+FCE+ICLE+ICNALE+Indian translated blog	45.4	50.1	45.7	49.9
Lang-8+FCE+ICLE+ICNALE+News+Tweets	45.2	57.5	45.5	55.2
Lang-8+FCE+ICLE+ICNALE+News+Tweets $r = 0.1$	44.9	**58.2**	45.0	**58.2**

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

Training Set	Accuracy (%)	
	no BA	with BA
Indian news	50.0	54.0
Indian tweets	54.0	**56.0**
Indian blogs	51.5	**56.0**

References and Acknowledgements

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. *TOEFL11: A corpus of non-native English*. Technical report, ETS.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media*.

Serhiy Bykh and Detmar Meurers. 2012. Native language identification using recurring *n*-grams – investigating abstraction and domain dependence. In *Proceedings of the 24th International Conference on Computational Linguistics*.

Max Coltheart. 1980. *MRC Psycholinguistic Database User Manual*. Birkbeck College.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses Universitaires de Louvain.

Shin'ichiro Ishikawa, 2011. A new horizon in learner corpus studies: The aim of the ICNALE project, In Weir, George and Shin'ichiro Ishikawa, and Kornwipa Poonpon (editors), *Corpora and Language Technologies in Teaching, Learning and Research*, pages 3–11. University of Strathclyde Press, Glasgow, UK.

Ben Swanson and Eugene Charniak. 2012. Native language detection with tree substitution grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. Summary report on the first shared task on native language identification. In *Proceedings of the Eighth Workshop on Building Educational Applications Using NLP.*

Yukio Tono, Yuji Kawaguchi, and Makoto Minegishi, editors. 2012. *Developmental and Cross-linguistic Perspectives in Learner Corpus Research*. John Benjamins, Amsterdam/Philadelphia.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

This work was supported by the Natural Sciences and Engineering Research Council of Canada. Thanks to Bo Han and Paul Cook for help with the tweets, and Varada Kolhatkar for help with Indian newpapers