Using Other Learner Corpora in the 2013 NLI Shared Task

Julian Brooke
Department of Computer Science
University of Toronto
jbrooke@cs.toronto.edu

Graeme Hirst
Department of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

Our efforts in the 2013 NLI shared task focused on the potential benefits of external corpora. We show that including training data from multiple corpora is highly effective at robust, cross-corpus NLI (i.e. open-training task 1), particularly when some form of domain adaptation is also applied. This method can also be used to boost performance even when training data from the same corpus is available (i.e. open-training task 2). However, in the closed-training task, despite testing a number of new features, we did not see much improvement on a simple model based on earlier work.

1 Introduction

Our participation in the 2013 NLI shared task (Tetreault et al., 2013) follows on our recent work exploring cross-corpus evaluation, i.e. using distinct corpora for training and testing (Brooke and Hirst, 2011; Brooke and Hirst, 2012a; Brooke and Hirst, 2012b), an approach that is now becoming fairly standard alternative in relevant work (Bykh and Meurers, 2012; Tetreault et al., 2012; Swanson and Charniak, 2013). Our promotion of cross-corpus evaluation in NLI was partially motivated by serious issues with the most popular corpus for native language identification work up to now, the International Corpus of Learner English (Granger et al., 2009). The new TOEFL-11 (Blanchard et al., 2013) used for this NLI shared task addresses some of the problems with the ICLE (most glaringly, the fact that some topics in the ICLE appeared only in some L1 backgrounds), but, from the perspective of topic, proficiency, and particularly genre, it is necessarily limited in scope (perhaps even more so than the ICLE); in short, it addresses only a small portion of the space of learner texts. Our interest, then, continues to be in robust models for NLI that are not restricted to utility in a particular corpus, and in our participation in this task we have focused our efforts on the open-training tasks which allow the use of corpora beyond the TOEFL-11. Since participation in these tasks was low relative to the closed-training task, fewer papers will address them, making our emphasis here all the more relevant.

The models built for all of three of the tasks are extensions of the model used in our recent work (Brooke and Hirst, 2012b); we will discuss the aspects of this model common to all tasks in Section 2. Section 3 is a brief review of our methodology and results in the closed-training task, which was focused exclusively on testing features (both new and old); we found almost nothing that improved on our best feature set from previous work, and most features actually hurt performance. In Section 4, we discuss the corpora we used for the open-training tasks, some of which we collected and/or have not been applied to NLI before. Our approach to the open-training task 2 using these corpora is presented in Section 5. In Section 6, we discuss how we used domain adaption methods and our various external corpora to create the (winning) model for the open-training task 1, which did not permit usage of the TOEFL-11; we also present some post hoc testing (now that TOEFL-11 is no longer off limits). In Section 7 we offer conclusions.

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 188–196,
Atlanta, Georgia, June 13 2013.

2 Basic Model

In our recent work on cross-corpus NLI (Brooke and Hirst, 2012b), we tested a number of classifier and feature options, and most of our choices there are carried over to this work. In particular, we use the Liblinear SVM 1va (one versus all) classifier (Fan et al., 2008). Using the TOEFL-11 corpus, we briefly tested the other options explored in that paper (including SVM 1v1) as well as the logistic regression classifier included in Liblinear, and found that the SVM 1va classifier was still preferred (with our best feature set, see below), though the differences involved were marginal. Although small variations in the choice of C parameter within the SVM model did occasionally produce benefits (here and in our previous work), these were not consistent, whereas the default value of 1 showed consistently near optimal results. We used a binary feature representation, and then feature vectors were normalized to the unit circle. With respect to feature selection, our earlier work used a frequency cutoff of 5 for all features; we continue to use frequency cutoffs here; other common feature selection methods (e.g. use of information gain) were ineffective in our previous work, so we did not explore them in detail here.

With regards to the features themselves, our earlier work tested a fairly standard collection of distributional features, including function words, word *n*-grams (up to bigram), POS *n*-grams (up to trigram), character *n*-grams (up to trigram), dependencies, context-free productions, and 'mixed' POS/function n-grams (up to trigram), i.e. *n*-grams with all lexical words replaced with part of speech. Most of these had appeared in previous NLI work (Koppel et al., 2005; Wong and Dras, 2011; Wong et al., 2012), though until recently word *n*-grams had been avoided because of ICLE topic bias. Our best model used only two of these features, word *n*-grams and the mixed POS/function *n*-grams. This was our starting point for the present work. The Stanford parser (Klein and Manning, 2003) was used for POS tagging and parsing.

Obviously, the training set used varies throughout the paper, and other differences in specific models built for each task will be mentioned as they become relevant. For evaluation here, we primarily use the test set for NLI shared task, though we employ some other evaluation corpora, as appropriate. During the preparation for the shared task, we made our decisions regarding models for two tasks with TOEFL-11 training according to the results in two training/test sets (800 per language for training, 100 per language for testing) sampled from the released training data. Since our research was focused on cross-corpus evaluation, we never created mechanisms for cross-validation in our system, and in fact it creates practical difficulties for the open-training task 2, so we do not include cross-validated results here.

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

3 Closed-training Task

Our approach to the closed-training task primarily involved feature testing. Table 1 contains the results of testing our previously investigated features from Brooke and Hirst (2012b) in the TOEFL-11, pivoted around the best set (word *n*-grams + mixed POS/Function *n*-grams) from that earlier work.

Some of the features we rejected in our previous work also underperform here, in particular character and POS *n*-grams. In fact, character *n*-grams had a much more negative effect on performance here than they had previously. Dependencies are clearly a useful feature in the TOEFL-11, this is fully consistent with out initial testing. CFG productions offer a small benefit on top of our base feature set, but are not useful when dependencies are also included, so we discarded them. Thus, our feature set going forward consists of word *n*-grams, mixed POS/function *n*-grams, and dependencies.

Next, we evaluate our feature frequency cutoff using this feature set (Table 2). We used the rather high cutoff of 5 (for all features) in the previous work because of our much larger training set. We looked at

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**

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

higher values there, but for this task we focused on testing lower values.

Lowering our frequency cutoff is indeed beneficial, and we got our best result in the test set when we had no feature selection at all. This was not consistent with our preparatory testing, which showed some benefit to removing hapax legomena, though the difference was marginal. However, we did include a run with this option in our final submission, and so this last result represents our best performance on the closed-training task.

We tested several other feature options that were added to our system for this task. Inspired by Bykh and Meurers (2012), we first considered *n*-grams (up to trigrams) where at least one lexical word is abstracted to its POS, and at least one isn't (partial abstraction). Since dependencies were found to be a positive feature, we tried adding dependency chains, which combine two dependencies, i.e. three lexical words linked by two grammatical relations. We tested productions with wild cards, e.g. S $\rightarrow$ NP VP * matches any sentence production which starts with NP VP. Tree Substitution grammar fragments have been shown to be superior to CFG productions (Swanson and Charniak, 2012); we used raw Tree Substitution Grammar (TSG) fragments for the TOEFL-11[1] and tested a subset of those fragments which involved at least two levels of the grammar (i.e. those not already covered by *n*-grams or CFG productions).

Our final feature option requires slightly more explanation. Crossley and McNamara (2012) report that metrics associated with word concreteness, imagability, meaningfulness, and familiarity are useful for NLI; the metrics they use are derived from the MRC Psycholinguistic database (Coltheart, 1980), which assign values for each dimension to individual words. We used the scores in the MRC to get an average score for each dimension for each text, further normalized to the range 0–1; texts with no words in the dictionaries were assigned the average across the training set.

Table 3 indicates that all of these new features were, to varying degrees, a drag on our model. The strongly negative effect of the MRC lexicons is particularly surprising. We speculate that this might might be due partially to problems with combining a large number of binary features with a small number of continuous metrics directly in a single SVM. A meta-classifier might solve this problem, but we did not explore meta-classification for features here.

Finally, since that information was available to us, we tested creating sub-models segregated by topic and proficiency. The topic-segregated model consisted of 8 SVMs, one for each topic; accuracy of this model was quite low, only 67.3%. The proficiency-segregated model used two groups, high and low/medium (there were few low texts, so we did not think they would be sufficient by themselves for a viable model). Results were higher, 74.9%, but still well below the best unsegregated model.

4 External Corpora

In this section we review corpora which will be used for the open-training tasks in the next two sections. Including the TOEFL-11, there are at least six publicly available multi-L1 learner text corpora for NLI, with many of these corpora becoming available relatively recently. Below, we introduce each corpus in detail; a summary of the number of tokens from each L1 background for each of the corpora is in Table 4.

[1]We thank Ben Swanson for letting us use his TSG fragments.

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

Lang-8 Lang-8 is a website where language learners write journal entries in their L2 to be corrected by native speakers. We collected a large set of these entries, which we've shown to be useful for NLI (Brooke and Hirst, 2012b), despite the noisiness of the corpus (for instance, some entries directly mix L1 and L2). For this task we added more entries written since the first version was collected (58k on top of the existing 154k entries).[2] The corpus contains entries from all the L1 backgrounds in the TOEFL-11, though the amounts for Hindi and particularly Telugu are small. Since many of the entries are very short, as in our previous work we add entries of the same L1 together to reach a minimum size of 250 tokens.

ICLE Before 2011, nearly all work on NLI was done in the International Corpus of Learner English or ICLE (Granger et al., 2009), a collection of college student essays from 15 L1 backgrounds, 8 of which overlap with the 11 L1s in the TOEFL-11. Despite known issues that might cause problems (Brooke and Hirst, 2011), it is probably the closest match in terms of genre and writer proficiency to the TOEFL-11.

FCE What we call the FCE corpus is a small sample of the First Certificate in English portion of the Cambridge Learner Corpus, which was released for the purposes of essay scoring evaluation (Yannakoudakis et al., 2011); 16 different L1 backgrounds are represented, 9 of which overlap with the TOEFL-11. Each of the texts consists of two short answers in the form of a letter, a report, an article, or a short story. Relative to the other corpora, the actual amount of text in the FCE is small.

ICCI Like the ICLE and TOEFL-11, the International Corpus of Crosslinguistic Interlanguage (Tono et al., 2012) is also an essay corpus, though in contrast with other corpora it is focused on young learners, i.e. those in grade school. It includes both descriptive and argumentative essays on a number of topics. Only 4 of its L1s overlap with the TOEFL-11.

ICANLE The International Corpus Network of Asian Learners of English or ICANLE (Ishikawa, 2011) is a collection of essays from college students in 10 Asian countries; 3 of the L1s overlap with the TOEFL-11.[3] Even more so than the TOEFL-11, this corpus is strictly controlled for topic, it has only 2 topics (part-time jobs and smoking in restaurants).

One obvious problem with using the above corpora to classify L1s in the TOEFL-11 is the lack of Hindi and Telugu text, which we found were the two most easily confused L1s in the closed-

[2] We do not have permission to distribute the corpus directly; however, we can offer a list of URLs together with software which can be used to recreate the corpus.

[3] The ICANLE also contains 103K of Urdu text. Since Urdu and Hindi are mutually intelligible, this could be a good substitute for Hindi; we overlooked this possibility during our preparation for the task, unfortunately.

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

training task. We explored a few methods to get data to fill this gap. First, we downloaded two collections of English language Indian news articles, one from a Hindi newspaper, the *Hindustan Times*, and one from a Telugu newspaper, the *Andhra Jyothy*.[4] Second, we extracted a collection of English tweets from the WORLD twitter corpus (Han et al., 2012) that were geolocated in the Hindi and Telugu speaking areas; as with the Lang-8, these were combined to create texts of at least 250 tokens.[5] Our third Indian corpus consists of translations (by Google Translate) of Hindi and Telugu blogs from the ICWSM 2009 Spinn3r Dataset (Burton et al., 2009), which we used in other work on using L1 text for NLI (Brooke and Hirst, 2012a). The number of tokens in each of these corpora are given in Table 5.

5 Open-training Task 2

Our approach to open-training task 2 is based on the assumption that in many ways it is a direct extension of the closed-training task. For example, we directly use the best feature set from that task, with no further testing. Based on the results in our initial testing, we used a feature frequency cutoff of 2 during our testing for open-training task 2; for consistency, we continue with that cutoff in this section.

We first attempted to integrate information from other corpora by using a meta-classifier, as was successfully used for features by Tetreault et al. (2012). Briefly, classifiers were trained on each major external corpus (including only the L1s in the TOEFL-11), and then tested on the TOEFL-11 training set; TOEFL-11 training was accomplished using 10-fold crossvalidation (by modifying the code for Liblinear crossvalidation to output margins). With the TOEFL-11 as the training set, the SVM margins from each 1va classifier (across all L1s and all corpora) were used as the feature input to the meta-classifier (also an SVM). In addition to Liblinear, we also outputted this meta-classification problem to WEKA format (Witten and Frank, 2005), and tested a number of other classifier options not available in Liblinear (e.g. Naïve Bayes, decision trees, random forests). In addition to (continuous) margins, we also tested using the classification directly. Ultimately, we came to the conclusion were that any use of a meta-classifer came with a cost (a minimum 2–3% drop in performance) that could not be fully overcome with the additional information from our external corpora. The result using SVM classifiers, margin features, and an SVM meta-classifier was 78.5%, well below the TOEFL-11–only baseline.

The other approach to using these external corpora is to add the data directly to the TOEFL-11 data and train a single classifier. This is very straightforward; really the only variable is which corpora will be included. However, we need to introduce, at this point, a domain-adaptation technique from our most recent work (Brooke and Hirst, 2012b), bias adaption, which we used to greatly improve the accuracy of cross-corpus classification. Without getting into the algorithmic details, bias adaption involves changing the bias (constant) factor of a model until the output of the model in some dataset is balanced across classes (or otherwise fits the expected distribution); it partially addresses skewed results due to differences between training and testing corpora. In the previous work, we used a separate development set, but here we rely on the test set itself; since the technique is unsupervised, we do not need to know the classes. Table 6 shows model performance after adding various corpora to the training set (TOEFL-11 is always included), with and without bias adaption (BA).

Many of the differences in Table 6 are modest, but there are are few points to be made. First, there is a small improvement using either the Lang-8 or the ICLE as additional data. The ICCI, on the other hand, has a clearly negative effect, perhaps be-

[4] As with the Lang-8, we cannot distribute the corpus directly but would be happy to provide URLs and scraping software for those would like to build it themselves.

[5] We extracted India regions 07 and 36 for Hindi, and 02 and 25 for Telegu; We can provide a list of tweet ids for reconstructing the corpus if desired. Our thanks to Bo Han and Paul Cook for helping us get these tweets.

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

cause of the age or proficiency of the contributors to that corpus. Bias adaption seems to help when the (messy and highly unbalanced) Lang-8 is involved (consistent with our previous work), but it does not seem useful applied to other corpora, at least not in this setting.

Our second adaptation technique involves training data selection, which has been used, for instance in cross-domain parsing (Plank and van Noord, 2011). The method used here is very simple: we count the number of times each word appears in a document in our test data, rank the texts in our training data according to the sum of counts (in the test data) each word that appears in a training texts, and throw away a certain numbers of low-ranked texts. For example, if a training text consists solely of the two words *I agree*[6] and *I* appears in 1053 texts in the test set, and *agree* appears in 325, then the value for that text is 1378. This method simultaneously penalizes short texts, those texts with low lexical diversity, and texts that do not use the same words as our test set. We use a fixed cutoff, r, which refers to the proportion of training data that is thrown away for each L1 (allowing this to work independent of L1 was not effective). We tested this on this method in tandem with bias adaption on two corpus sets: The TOEFL-11 and the Lang-8, and all corpora except the ICCI. The results are in Table 7. The number in italics is the best run that we submitted.

Again, it is difficult to come to any firm conclusions when the differences are this small, but

[6]This is not a made-up example; there is actually a text in the TOEFL-11 corpus like this.

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

our best results involve all of the corpora (except the ICCI) and both adaptation techniques. Unfortunately, our initial testing suggested $r = 0.2$ was the better choice, so our official best result in this task (81.6%) is not the best result in this table. Performance clearly drops for $r > 0.2$. Nevertheless, nearly all the results in the table show clear improvement on our closed-training task model.

6 Open-training Task 1

The central challenge of open-training task 1 was that the TOEFL-11 was completely off-limits, even for testing. Therefore, a discussion of how we prepared for this task is very distinct from a post hoc analysis of the best method once we allowed ourselves access to the TOEFL-11; we separate the two here. We did use the feature set (and frequency cutoff) from the closed-training (and open-training 2) task; it was close enough to the feature set from our earlier work (using the Lang-8, ICLE, and FCE) that it did not seem like cheating to preserve it.

6.1 Method

Given our failure to create a meta-classifier in open-training task 2, we did not pursue that option here, focusing purely on adding corpora directly to a mixed training set. The central question was which corpora to add, and whether to use our domain-adaptation methods. Our experience with the ICCI in the open-training task 2 suggested that it might be worth leaving it (or perhaps other corpora) out, but

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**

could we come to that conclusion independently?

Our approach involved considering each external corpus as a test set, and seeing which other corpora were useful when included in the training set; corpora which were consistently useful would be included in the final set. Our original exploration involved looking at all of the corpora (as test sets), but it was haphazard; here, we present results just with the ICLE and the ICANLE, which are arguably the two closest corpora to the TOEFL-11 in terms of proficiency and genre. For this, we used a different selection of L1s, 12 for the ICLE, 7 for the ICANLE; all of these languages appeared in at least the Lang-8, and 2 of them (Chinese and Japanese) appeared in all corpora. Both sets were balanced by L1. Again, we report results with and without bias adaption. The results for the ICLE are in Table 8.

The clearest result in Table 8 is the consistently positive effect of bias adaption, at least 10 percentage points, which is line with our previous work. Adding both ICLE and ICNALE to the Lang-8 corpus gave a small boost in performance, but the effect of the ICCI was once again negative, as was the effect of our training set selection.

The ICNALE results in Table 9 support many of the conclusions that we reached in the ICLE (and other sets like the FCE and ICCI, which are not included here but gave similar results); the effect of bias adaption is even more pronounced. Two differences: the slightly positive effect of training data selection and the positive effect of the ICCI, the latter of which we saw nowhere else. We speculate that this might be due to that fact that although the ICNALE is a college-level corpus, it is a corpus of Asian-language native speakers. Our theory is that Europeans are, on average, more proficient users of English (this is supported by, for instance, the testing from Granger et al. (2009)), and that therefore the European component of the low-proficiency ICCI actually interferes with using high proficiency as a way of distinguishing European L1s, a problem which would obviously not extend to an Asian-L1-only corpus. This is an interesting result, but we will not explore it further here. In any case, it would lead us to predict that including ICCI data would be a bad idea for TOEFL-11 testing.

Since we did not have any way to evaluate our Indian corpora (i.e. the news, twitter, and translated blogs from Section 4) without using the TOEFL-11, we instead took advantage of the option to submit multiple runs, submitting runs which use each of the corpora, and combining the blogs and news.

6.2 Post Hoc Analysis

With the TOEFL-11 data now visible to us, we first ask whether our specially collected Indian corpora can distinguish texts in the ICCI. The test set used in Table 10 contains only Hindi and Telugu texts. The results are quite modest (the guessing baseline is 50%), but suggest that all three corpora contain some information that distinguish Hindi and Telugu, particularly if bias adaption is used.

The results for a selection of models on the full set of TOEFL-11 languages is presented in Table 11. Since ours was the best-performing model in this task, we include results for both the TOEFL-11 training (including development set) and test set, to facilitate future comparison. Again, there is little doubt that bias adaption is of huge benefit, though in fact our results in the Lang-8 alone, without bias

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**

adaption, would have been enough to take first place in this task. Adding other corpora, including the Indian corpora but not the ICCI, did consistently improve performance, as suggested by our testing in other corpora. Although the translated blog data was useful in distinguishing Hindi from Telugu alone, it had an unpredictable effect in the main task, lowering bias-adapted performance. Training set selection does seem to have a small positive effect, though we did not see this consistently in our original testing.

7 Conclusion

Our efforts in the 2013 NLI shared task focused on the potential benefits of external corpora. We have shown here that including training data from multiple corpora is effective at creating good cross-corpus NLI systems, particularly when domain adaptation, i.e. bias adaption or training set selection, is also applied; we were the highest-performing group in open-training task 1 by a large margin. This approach can also be applied to improve performance even when training data from the same corpus is available, as in open-training task 2. However, in the closed-training task, despite testing a number of new features, we did not see much improvement on our simple model based on earlier work. Other teams clearly did find some ways to improve on this straightforward approach, and we hope to see to what extent those improvements are generalizable across different NLI corpora.

Acknowledgements

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada.

References

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

Julian Brooke and Graeme Hirst. 2011. Native language detection with 'cheap' learner corpora. Presented at the 2011 Learner Corpus Research Conference. Published in Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier, editors, (2013) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use - Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain.

Julian Brooke and Graeme Hirst. 2012a. Measuring interlanguage: Native language identification with L1-influence metrics. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, pages 779–784, Istanbul, Turkey.

Julian Brooke and Graeme Hirst. 2012b. Robust, lexicalized native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

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 (ICWSM 2009)*, San Jose, CA.

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 (COLING '12)*.

Max Coltheart. 1980. *MRC Psycholinguistic Database User Manual: Version 1*. Birkbeck College.

Scott A. Crossley and Danielle S. McNamara. 2012. Detecting the first language of second language writers using automated indicies of cohesion, lexical sophistication, syntactic complexity and conceptual knowledge. In Scott Jarvis and Scott A. Crossley, editors, *Approaching Language Transfer through Text Classification: Explorations in the Detection-based Approach*. Multilingual Matters.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

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

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

Shin'ichiro Ishikawa, 2011. *A new horizon in learner corpus studies: The aim of the ICNALE project*, pages 3–11. University of Strathclyde Press, Glasgow, UK.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pages 624–628, Chicago, Illinois, USA.

Barbara Plank and Gertjan van Noord. 2011. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1566–1576, Portland, Oregon, USA, June.

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 (ACL '12)*, pages 193–197, Jeju, Korea.

Ben Swanson and Eugene Charniak. 2013. Extracting the native language signal for second language acquisition. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '13)*.

Joel Tetreault, Daniel Blanchard, Aoife Cahill, and Martin Chodorow. 2012. Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING '12)*.

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*, Atlanta, GA, USA, June. Association for Computational Linguistics.

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

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, pages 1600–1610, Edinburgh, Scotland, UK.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2012. Exploring adaptor grammars for native language identification. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*, Jeju, Korea.

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*, pages 180–189, Portland, Oregon.