A Report on the First Native Language Identification Shared Task

Joel Tetreault*, **Daniel Blanchard**† **and Aoife Cahill**†

* Nuance Communications, Inc., 1198 E. Arques Ave, Sunnyvale, CA 94085, USA
Joel.Tetreault@nuance.com
† Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA
{dblanchard, acahill}@ets.org

Abstract

Native Language Identification, or NLI, is the task of automatically classifying the L1 of a writer based solely on his or her essay written in another language. This problem area has seen a spike in interest in recent years as it can have an impact on educational applications tailored towards non-native speakers of a language, as well as authorship profiling. While there has been a growing body of work in NLI, it has been difficult to compare methodologies because of the different approaches to pre-processing the data, different sets of languages identified, and different splits of the data used. In this shared task, the first ever for Native Language Identification, we sought to address the above issues by providing a large corpus designed specifically for NLI, in addition to providing an environment for systems to be directly compared. In this paper, we report the results of the shared task. A total of 29 teams from around the world competed across three different sub-tasks.

1 Introduction

One quickly growing subfield in NLP is the task of identifying the native language (L1) of a writer based solely on a sample of their writing in another language. The task is framed as a classification problem where the set of L1s is known *a priori*. Most work has focused on identifying the native language of writers learning English as a second language. To date this topic has motivated several papers and research projects.

Native Language Identification (NLI) can be useful for a number of applications. NLI can be used in educational settings to provide more targeted feedback to language learners about their errors. It is well known that speakers of different languages make different kinds of errors when learning a language (Swan and Smith, 2001). A writing tutor system which can detect the native language of the learner will be able to tailor the feedback about the error and contrast it with common properties of the learner's language. In addition, native language is often used as a feature that goes into authorship profiling (Estival et al., 2007), which is frequently used in forensic linguistics.

Despite the growing interest in this field, development has been encumbered by two issues. First is the issue of data. Evaluating an NLI system requires a corpus containing texts in a language other than the native language of the writer. Because of a scarcity of such corpora, most work has used the International Corpus of Learner English (ICLEv2) (Granger et al., 2009) for training and evaluation since it contains several hundred essays written by college-level English language learners. However, this corpus is quite small for training and testing statistical systems which makes it difficult to tell whether the systems that are developed can scale well to larger data sets or to different domains.

Since the ICLE corpus was not designed with the task of NLI in mind, the usability of the corpus for this task is further compromised by idiosyncrasies in the data such as topic bias (as shown by Brooke and Hirst (2011)) and the occurrence of characters which only appear in essays written by speakers of certain languages (Tetreault et al., 2012). As a result, it is hard to draw conclusions about which features

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 48–57,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

actually perform best. The second issue is that there has been little consistency in the field in the use of cross-validation, the number of L1s, and which L1s are used. As a result, comparing one approach to another has been extremely difficult.

The first Shared Task in Native Language Identification is intended to better unify this community and help the field progress. The Shared Task addresses the two deficiencies above by first using a new corpus (TOEF11, discussed in Section 3) that is larger than the ICLE and designed specifically for the task of NLI and second, by providing a common set of L1s and evaluation standards that everyone will use for this competition, thus facilitating direct comparison of approaches. In this report we describe the methods most participants used, the data they evaluated their systems on, the three sub-tasks involved, the results achieved by the different teams, and some suggestions and ideas about what we can do for the next iteration of the NLI shared task.

In the following section, we provide a summary of the prior work in Native Language Identification. Next, in Section 3 we describe the TOEFL11 corpus used for training, development and testing in this shared task. Section 4 describes the three sub-tasks of the NLI Shared Task as well as a review of the timeline. Section 5 lists the 29 teams that participated in the shared task, and introduce abbreviations that will be used throughout this paper. Sections 6 and 7 describe the results of the shared task and a separate post shared task evaluation where we asked teams to evaluate their system using cross-validation on a combination of the training and development data. In Section 8 we provide a high-level view of the common features and machine learning methods teams tended to use. Finally, we offer conclusions and ideas for future instantiations of the shared task in Section 9.

2 Related Work

In this section, we provide an overview of some of the common approaches used for NLI prior to this shared task. While a comprehensive review is outside the scope of this paper, we have compiled a bibliography of related work in the field. It can be downloaded from the NLI Shared Task website.[1]

To date, nearly all approaches have treated the task of NLI as a supervised classification problem where statistical models are trained on data from the different L1s. The work of Koppel et al. (2005) was the first in the field and they explored a multitude of features, many of which are employed in several of the systems in the shared tasks. These features included character and POS n-grams, content and function words, as well as spelling and grammatical errors (since language learners have tendencies to make certain errors based on their L1 (Swan and Smith, 2001)). An SVM model was trained on these features extracted from a subsection of the ICLE corpus consisting of 5 L1s.

N-gram features (word, character and POS) have figured prominently in prior work. Not only are they easy to compute, but they can be quite predictive. However, there are many variations on the features. Past reseach efforts have explored different n-gram windows (though most tend to focus on unigrams and bigrams), different thresholds for how many n-grams to include as well as whether to encode the feature as binary (presence or absence of the particular n-gram) or as a normalized count.

The inclusion of syntactic features has been a focus in recent work. Wong and Dras (2011) explored the use of production rules from two parsers and Swanson and Charniak (2012) explored the use of Tree Substitution Grammars (TSGs). Tetreault et al. (2012) also investigated the use of TSGs as well as dependency features extracted from the Stanford parser.

Other approaches to NLI have included the use of Latent Dirichlet Analysis to cluster features (Wong et al., 2011), adaptor grammars (Wong et al., 2012), and language models (Tetreault et al., 2012). Additionally, there has been research into the effects of training and testing on different corpora (Brooke and Hirst, 2011).

Much of the aforementioned work takes the perspective of optimizing for the task of Native Language Identification, that is, what is the best way of modeling the problem to get the highest system accuracy? The problem of Native Language Identifica-

[1] http://nlisharedtask2013.org/bibliography-of-related-work-in-nli

tion is also of interest to researchers in Second Language Acquisition where they seek to explain syntactic transfer in learner language (Jarvis and Crossley, 2012).

3 Data

The dataset for the task was the new TOEFL11 corpus (Blanchard et al., 2013). TOEFL11 consists of essays written during a high-stakes college-entrance test, the Test of English as a Foreign Language (TOEFL®). The corpus contains 1,100 essays per language sampled as evenly as possible from 8 prompts (i.e., topics) along with score levels (low/medium/high) for each essay. The 11 native languages covered by our corpus are: Arabic (ARA), Chinese (CHI), French (FRE), German (GER), Hindi (HIN), Italian (ITA), Japanese (JAP), Korean (KOR), Spanish (SPA), Telugu (TEL), and Turkish (TUR).

The TOEFL11 corpus was designed specifically to support the task of native language identification. Because all of the essays were collected through ETS's operational test delivery system for the TOEFL® test, the encoding and storage of all texts in the corpus is consistent. Furthermore, the sampling of essays was designed to ensure approximately equal representation of native languages across topics, insofar as this was possible.

For the shared task, the corpus was split into three sets: training (TOEFL11-TRAIN), development (TOEFL11-DEV), and test (TOEFL11-TEST). The train corpus consisted of 900 essays per L1, the development set consisted of 100 essays per L1, and the test set consisted of another 100 essays per L1. Although the overall TOEFL11 corpus was sampled as evenly as possible with regard to language and prompts, the distribution for each language is not exactly the same in the training, development and test sets (see Tables 1a, 1b, and 1c). In fact, the distribution is much closer between the training and test sets, as there are several languages for which there are no essays for a given prompt in the development set, whereas there are none in the training set, and only one, Italian, for the test set.

It should be noted that in the first instantiation of the corpus, presented in Tetreault et al. (2012), we used TOEFL11 to denote the body of data consisting of TOEFL11-TRAIN and TOEFL11-DEV. However, in this shared task, we added 1,100 sentences for a test set and thus use the term TOEFL11 to now denote the corpus consisting of the TRAIN, DEV and TEST sets. We expect the corpus to be released through the the Linguistic Data Consortium in 2013.

4 NLI Shared Task Description

The shared task consisted of three sub-tasks. For each task, the test set was TOEFL11-TEST and only the type of training data varied from task to task.

- **Closed-Training**: The first and main task was the 11-way classification task using only the TOEFL11-TRAIN and optionally TOEFL11-DEV for training.
- **Open-Training-1**: The second task allowed the use of any amount or type of training data (as is done by Brooke and Hirst (2011)) *excluding* any data from the TOEFL11, but still evaluated on TOEFL11-TEST.
- **Open-Training-2**: The third task allowed the use of TOEFL11-TRAIN and TOEFL11-DEV combined with any other additional data. This most closely reflects a real-world scenario.

Additionally, each team could submit up to 5 different systems per task. This allowed a team to experiment with different variations of their core system.

The training data was released on January 14, with the development data and evaluation script released almost one month later on February 12. The train and dev data contained an index file with the L1 for each essay in those sets. The previously unseen and unlabeled test data was released on March 11 and teams had 8 days to submit their system predictions. The predictions for each system were encoded in a CSV file, where each line contained the file ID of a file in TOEFL11-TEST and the corresponding L1 prediction made by the system. Each CSV file was emailed to the NLI organizers and then evaluated against the gold standard.

5 Teams

In total, 29 teams competed in the shared task competition, with 24 teams electing to write papers describing their system(s). The list of participating

Lang.	P1	P2	P3	P4	P5	P6	P7	P8
ARA	113	113	113	112	112	113	112	112
CHI	113	113	113	112	112	113	112	112
FRE	128	128	76	127	127	60	127	127
GER	125	125	125	125	125	26	125	124
HIN	132	132	132	71	132	38	132	131
ITA	142	70	122	141	141	12	141	131
JAP	108	114	113	113	113	113	113	113
KOR	113	113	113	112	112	113	112	112
SPA	124	120	38	124	123	124	124	123
TEL	139	139	139	41	139	26	139	138
TUR	132	132	72	132	132	37	132	131
Total	1369	1299	1156	1210	1368	775	1369	1354

(a) Training Set

Lang.	P1	P2	P3	P4	P5	P6	P7	P8
ARA	12	13	13	13	14	7	14	14
CHI	14	14	0	15	15	14	13	15
FRE	17	18	0	14	19	0	13	19
GER	15	15	16	10	13	0	15	16
HIN	16	17	17	0	17	0	16	17
ITA	18	0	0	30	31	0	21	0
JAP	0	14	15	14	15	14	14	14
KOR	15	8	15	2	13	15	16	16
SPA	7	0	0	21	7	21	21	23
TEL	16	17	17	0	17	0	16	17
TUR	22	4	0	22	7	0	22	23
Total	152	120	93	141	168	71	181	174

(b) Dev Set

Lang.	P1	P2	P3	P4	P5	P6	P7	P8
ARA	13	11	12	14	10	13	12	15
CHI	13	14	13	13	7	14	14	12
FRE	13	14	11	15	14	8	11	14
GER	15	14	16	16	12	2	12	13
HIN	13	13	14	15	7	15	10	13
ITA	13	19	16	16	15	0	11	10
JAP	8	14	12	11	10	15	14	16
KOR	12	12	8	14	12	14	13	15
SPA	10	13	16	14	4	12	15	16
TEL	10	10	11	14	13	15	11	16
TUR	15	9	18	16	8	6	13	15
Total	135	143	147	158	112	114	136	155

(c) Test Set

Table 1: Number of essays per language per prompt in each data set

teams, along with their abbreviations, can be found in Table 2.

6 Shared Task Results

This section summarizes the results of the shared task. For each sub-task, we have tables listing the

Team Name	**Abbreviation**
Bobicev	BOB
Chonger	CHO
CMU-Haifa	HAI
Cologne-Nijmegen	CN
CoRAL Lab @ UAB	COR
CUNI (Charles University)	CUN
cywu	CYW
dartmouth	DAR
eurac	EUR
HAUTCS	HAU
ItaliaNLP	ITA
Jarvis	JAR
kyle, crossley, dai, mcnamara	KYL
LIMSI	LIM
LTRC IIIT Hyderabad	HYD
Michigan	MIC
MITRE “Carnie”	CAR
MQ	MQ
NAIST	NAI
NRC	NRC
Oslo NLI	OSL
Toronto	TOR
Tuebingen	TUE
Ualberta	UAB
UKP	UKP
Unibuc	BUC
UNT	UNT
UTD	UTD
VTEX	VTX

Table 2: Participating Teams and Team Abbreviations

top submission for each team and its performance by overall accuracy and by L1.[2]

Table 3 shows results for the Closed sub-task where teams developed systems that were trained solely on TOEFL11-TRAIN and TOEFL11-DEV. This was the most popular sub-task with 29 teams competing and 116 submissions in total for the sub-task. Most teams opted to submit 4 or 5 runs.

The Open sub-tasks had far fewer submissions. Table 4 shows results for the Open-1 sub-task where teams could train systems using any training data *excluding* TOEFL11-TRAIN and TOEFL11-DEV. Three teams competed in this sub-task for a total of 13 submissions. Table 5 shows the results for the third sub-task “Open-2”. Four teams competed in this task for a total of 15 submissions.

The challenge for those competing in the Open tasks was finding enough non-TOEFL11 data for each L1 to train a classifier. External corpora commonly used in the competition included the:

- **ICLE**: which covered all L1s except for Arabic, Hindi and Telugu;

- **FCE: First Certificate in English Corpus** (Yannakoudakis et al., 2011): a collection of essay written for an English assessment exam, which covered all L1s except for Arabic, Hindi and Telugu

- **ICNALE: International Corpus Network of Asian Learners of English** (Ishikawa, 2011): a collection of essays written by Chinese, Japanese and Korean learners of English along with 7 other L1s with Asian backgrounds.

- **Lang8: http://www.lang8.com**: a social networking service where users write in the language they are learning, and get corrections from users who are native speakers of that language. Shared Task participants such as NAI and TOR scraped the website for all writng samples from English language learners. All of the L1s in the shared task are represented on the site, though the Asian L1s dominate.

The most challenging L1s to find data for seemed to be Hindi and Telugu. TUE used essays written by Pakastani students in the ICNALE corpus to substitute for Hindi. For Telugu, they scraped material from bilingual blogs (English-Telugu) as well as other material for the web. TOR created corpora for Telugu and Hindi by scraping news articles, tweets which were geolocated in the Hindi and Telugu speaking areas, and translations of Hindi and Telugu blogs using Google Translate.

We caution directly comparing the results of the Closed sub-task to the Open ones. In the Open-1 sub-task most teams had smaller training sets than used in the Closed competition which automatically puts them at a disadvantage, and in some cases there

[2] For those interested in the results of all submissions, please contact the authors.

			L1 F-Score										
Team Name	**Run**	**Overall Acc.**	**ARA**	**CHI**	**FRE**	**GER**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**
JAR	2	0.836	0.785	0.856	0.860	0.893	0.775	0.905	0.854	0.813	0.798	0.802	0.854
OSL	2	0.834	0.816	0.850	0.874	0.912	0.792	0.873	0.828	0.806	0.783	0.792	0.840
BUC	5	0.827	0.840	0.866	0.853	0.931	0.736	0.873	0.851	0.812	0.779	0.760	0.796
CAR	2	0.826	0.859	0.847	0.810	0.921	0.762	0.877	0.825	0.827	0.768	0.802	0.790
TUE	1	0.822	0.810	0.853	0.806	0.897	0.768	0.883	0.842	0.776	0.772	0.824	0.812
NRC	4	0.818	0.804	0.845	0.848	0.916	0.745	0.903	0.818	0.790	0.788	0.755	0.790
HAI	1	0.815	0.804	0.842	0.835	0.903	0.759	0.845	0.825	0.806	0.776	0.789	0.784
CN	2	0.814	0.778	0.845	0.848	0.882	0.744	0.857	0.812	0.779	0.787	0.784	0.827
NAI	1	0.811	0.814	0.829	0.828	0.876	0.755	0.864	0.806	0.789	0.757	0.793	0.802
UTD	2	0.809	0.778	0.846	0.832	0.892	0.731	0.866	0.846	0.819	0.715	0.784	0.784
UAB	3	0.803	0.820	0.804	0.822	0.905	0.724	0.850	0.811	0.736	0.777	0.792	0.786
TOR	1	0.802	0.754	0.827	0.827	0.878	0.722	0.850	0.820	0.808	0.747	0.784	0.798
MQ	4	0.801	0.800	0.828	0.789	0.885	0.738	0.863	0.826	0.780	0.703	0.782	0.802
CYW	1	0.797	0.769	0.839	0.782	0.833	0.755	0.842	0.815	0.770	0.741	0.828	0.788
DAR	2	0.781	0.761	0.806	0.812	0.870	0.706	0.846	0.788	0.776	0.730	0.723	0.767
ITA	1	0.779	0.738	0.775	0.832	0.873	0.711	0.860	0.788	0.742	0.708	0.762	0.780
CHO	1	0.775	0.764	0.835	0.798	0.888	0.721	0.816	0.783	0.670	0.688	0.786	0.758
HAU	1	0.773	0.731	0.820	0.806	0.897	0.686	0.830	0.832	0.763	0.703	0.702	0.736
LIM	4	0.756	0.737	0.760	0.788	0.886	0.654	0.808	0.775	0.756	0.712	0.701	0.745
COR	5	0.748	0.704	0.806	0.783	0.898	0.670	0.738	0.794	0.739	0.616	0.730	0.741
HYD	1	0.744	0.680	0.778	0.748	0.839	0.693	0.788	0.781	0.735	0.613	0.770	0.754
CUN	1	0.725	0.696	0.743	0.737	0.830	0.714	0.838	0.676	0.670	0.680	0.697	0.684
UNT	3	0.645	0.667	0.682	0.635	0.746	0.558	0.687	0.676	0.620	0.539	0.667	0.609
BOB	4	0.625	0.513	0.684	0.638	0.751	0.612	0.706	0.647	0.549	0.495	0.621	0.608
KYL	1	0.590	0.589	0.603	0.643	0.634	0.554	0.663	0.627	0.569	0.450	0.649	0.507
UKP	2	0.583	0.592	0.560	0.624	0.653	0.558	0.616	0.631	0.565	0.456	0.656	0.489
MIC	3	0.430	0.419	0.386	0.411	0.519	0.407	0.488	0.422	0.384	0.400	0.500	0.396
EUR	1	0.386	0.500	0.390	0.277	0.379	0.487	0.522	0.441	0.352	0.281	0.438	0.261
VTX	5	0.319	0.367	0.298	0.179	0.297	0.159	0.435	0.340	0.370	0.201	0.410	0.230

Table 3: Results for closed task

			L1 F-Score										
Team Name	**Run**	**Overall Acc.**	**ARA**	**CHI**	**FRE**	**GER**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**
TOR	5	0.565	0.410	0.776	0.692	0.754	0.277	0.680	0.660	0.650	0.653	0.190	0.468
TUE	2	0.385	0.114	0.502	0.420	0.430	0.167	0.611	0.485	0.348	0.385	0.236	0.314
NAI	2	0.356	0.329	0.450	0.331	0.423	0.066	0.511	0.426	0.481	0.314	0.000	0.207

Table 4: Results for open-1 task

			L1 F-Score										
Team Name	**Run**	**Overall Acc.**	**ARA**	**CHI**	**FRE**	**GER**	**HIN**	**ITA**	**JPN**	**KOR**	**SPA**	**TEL**	**TUR**
TUE	1	0.835	0.798	0.876	0.844	0.883	0.777	0.883	0.836	0.794	0.846	0.826	0.818
TOR	4	0.816	0.770	0.861	0.840	0.900	0.704	0.860	0.834	0.800	0.816	0.804	0.790
HYD	1	0.741	0.677	0.782	0.755	0.829	0.693	0.784	0.777	0.728	0.613	0.766	0.744
NAI	3	0.703	0.676	0.695	0.708	0.846	0.618	0.830	0.677	0.610	0.663	0.726	0.688

Table 5: Results for open-2 task

was a mismatch in the genre of corpora (for example, tweets by Telugu speakers are different in composition than essays written by Telugu speakers). TUE and TOR were the only two teams to participate in all three sub-tasks, and their Open-2 systems outperformed their respective best systems in the Closed and Open-1 sub-tasks. This suggests, unsurprisingly, that adding more data can benefit NLI, though quality and genre of data are also important factors.

7 Cross Validation Results

Upon completion of the competition, we asked the participants to perform 10-fold cross-validation on a data set consisting of the union of TOEFL11-TRAIN and TOEFL11-DEV. This was the same set of data used in the first work to use any of the TOEFL11 data (Tetreault et al., 2012), and would allow another point of comparison for future NLI work. For direct comparison with Tetreault et al. (2012), we provided the exact folds used in that work.

The results of the 10-fold cross-validation are shown in Table 6. Two teams had systems that performed at 84.5 or better, which is just slightly higher than the best team performance on the TOEFL11-TEST data. In general, systems that performed well in the main competition also performed similarly (in terms of performance and ranking) in the cross-validation experiment. Please note that we report results as they are reported in the respective papers, rounding to just one decimal place where possible.

Team	Accuracy
CN	84.6
JAR	84.5
OSL	83.9
BUC	82.6
MQ	82.5
TUE	82.4
CAR	82.2
NAI	82.1
Tetreault et al. (2012)	80.9
HAU	79.9
LIM	75.9
CUN	74.2
UNT	63.8
MIC	63

Table 6: Results for 10-fold cross-validation on TOEFL11-TRAIN + TOEFL11-DEV

8 Discussion of Approaches

With so many teams competing in the shared task competition, we investigated whether there were any commonalities in learning methods or features between the teams. In this section, we provide a coarse grained summary of the common machine learning methods teams employed as well as some of the common features. Our summary is based on the information provided in the 24 team reports.

While there are many machine learning algorithms to choose from, the overwhelming majority of teams used Support Vector Machines. This may not be surprising given that most prior work has also used SVMs. Tetreault et al. (2012) showed that one could achieve even higher performance on the NLI task using ensemble methods for combining classifiers. Four teams also experimented with different ways of using ensemble methods. Three teams used Maximum Entropy methods for their modeling. Finally, there were a few other teams that tried different methods such as Discriminant Function Analysis and K-Nearest Neighbors. Possibly the most distinct method employed was that of string kernels by the BUC team (who placed third in the closed competition). This method only used character level features. A summary of the machine learning methods is shown in Table 7.

A summary of the common features used across teams is shown in Table 8. It should be noted that the table does not detail the nuanced differences in how the features are realized. For example, in the case of n-grams, some teams used only the top k most frequently n-grams while others used all of the n-grams available. If interested in more information about the particulars of a system and its feature, we recommend reading the team's summary report.

The most common features were word, character and POS n-gram features. Most teams used n-grams ranging from unigrams to trigrams, in line with prior literature. However several teams used higher-order n-grams. In fact, four of the top five teams (JAR, OSL, CAR, TUE) generally used at least 4-grams,

Machine Learning	Teams
SVM	CN, UNT, MQ, JAR, TOR, ITA, CUN, TUE, COR, NRC, HAU, MIC, CAR
MaxEnt / logistic regression	LIM, HAI, CAR
Ensemble	MQ, ITA, NRC, CAR
Discriminant Function Analysis	KYL
String Kernels / LRD	BUC
PPM	BOB
k-NN	VTX

Table 7: Machine Learning algorithms used in Shared Task

and some, such as OSL and JAR, went as high 7 and 9 respectively in terms of character n-grams.

Syntactic features, which were first evaluated in Wong and Dras (2011) and Swanson and Charniak (2012) were used by six teams in the competition, with most using dependency parses in different ways. Interestingly, while Wong and Dras (2011) showed some of the highest performance scores on the ICLE corpus using parse features, only two of the six teams which used them placed in the top ten in the Closed sub-task.

Spelling features were championed by Koppel et al. (2005) and in subsequent NLI work, however only three teams in the competition used them.

There were several novel features that teams tried. For example, several teams tried skip n-grams, as well as length of words, sentences and documents; LIM experimented with machine translation; CUN had different features based on the relative frequencies of the POS and lemma of a word; HAI tried several new features based on passives and context function; and the TUE team tried a battery of syntactic features as well as text complexity measures.

9 Summary

We consider the first edition of the shared task a success as we had 29 teams competing, which we consider a large number for any shared task. Also of note is that the task brought together researchers not only from the Computational Linguistics community, but also those from other linguistics fields such as Second Language Acquisition.

We were also delighted to see many teams build on prior work but also try novel approaches. It is our hope that finally having an evaluation on a common data set will allow researchers to learn from each other on what works well and what does not, and thus the field can progress more rapidly. The evaluation scripts are publicly available and we expect that the data will become available through the Linguistic Data Consortium in 2013.

For future editions of the NLI shared task, we think it would be interesting to expand the scope of NLI from identifying the L1 of student essays to be able to identify the L1 of any piece of writing. The ICLE and TOEFL11 corpora are both collections of academic writing and thus it may be the case that certain features or methodologies generalize better to other writing genres and domains. For those interested in robust NLI approaches, please refer to the TOR team shared task report as well as Brooke and Hirst (2012).

In addition, since the TOEFL11 data contains proficiency level one could include an evaluation by proficiency level as language learners make different types of errors and may even have stylistic differences in their writing as their proficiency progresses.

Finally, while this may be in the periphery of the scope of an NLI shared task, one interesting evaluation is to see how well human raters can fare on this task. This would of course involve knowledgeable language instructors who have years of experience in teaching students from different L1s. Our thinking is that NLI might be one task where computers would outperform human annotators.

Acknowledgments

We would like to thank Derrick Higgins and members of Educational Testing Service for assisting us in making the TOEFL11 essays available for this shared task. We would also like to thank Patrick Houghton for assisting the shared task organizers.

Feature	Type	Teams
Word N-Grams	1	CN, UNT, JAR, TOR, KYL, ITA, CUN, BOB, OSL, TUE, UAB, CYW, NAI, NRC, MIC, CAR
	2	CN, UNT, JAR, TOR, KYL, ITA, CUN, BOB, OSL, TUE, COR, UAB, CYW, NAI, NRC, HAU, MIC, CAR
	3	UNT, MQ, JAR, KYL, CUN, COR, HAU, MIC, CAR
	4	JAR, KYL, CAR
	5	CAR
POS N-grams	1	CN, UNT, JAR, TOR, ITA, LIM, CUN, BOB, TUE, HAI, CAR
	2	CN, UNT, JAR, TOR, ITA, LIM, CUN, BOB, TUE, COR, HAI, NAI, NRC, MIC, CAR
	3	CN, UNT, JAR, TOR, LIM, CUN, TUE, COR, HAI, NAI, NRC, CAR
	4	CN, JAR, TUE, HAI, NRC, CAR
	5	TUE, CAR
Character N-Grams	1	CN, UNT, MQ, JAR, TOR, LIM, BOB, OSL, HAI, CAR
	2	CN, UNT, MQ, JAR, TOR, ITA, LIM, BOB, OSL, COR, HAI, NAI, HAU, MIC, CAR
	3	CN, UNT, MQ, JAR, TOR, LIM, BOB, OSL, VTX, COR, HAI, NAI, NRC, HAU, MIC, CAR
	4	CN, JAR, LIM, BOB, OSL, HAI, HAU, MIC, CAR
	5	CN, JAR, BOB, OSL, HAU, CAR
	6	CN, JAR, OSL,
	7	JAR, OSL
	8-9	JAR
Function N-Grams		MQ, UAB
Syntactic Features	Dependencies	MQ, TOR, ITA, TUE, NAI, NRC
	TSG	MQ, TOR, NAI,
	CF Productions	TOR,
	Adaptor Grammars	MQ
Spelling Features		LIM,CN, HAI

Table 8: Common Features used in Shared Task

In addition, thanks goes to the BEA8 Organizers (Joel Tetreault, Jill Burstein and Claudia Leacock) for hosting the shared task with their workshop. Finally, we would like to thank all the teams for participating in this first shared task and making it a success. Their feedback, patience and enthusiasm made organizing this shared task a great experience.

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. In *Conference of Learner Corpus Research (LCR2011)*, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.

Julian Brooke and Graeme Hirst. 2012. Robust, Lexicalized Native Language Identification. In *Proceedings of COLING 2012*, pages 391–408, Mumbai, India, December. The COLING 2012 Organizing Committee.

Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 263–272, Melbourne, Australia.

Sylviane Granger, Estelle Dagneaux, and Fanny Meunier. 2009. *The International Corpus of Learner English: Handbook and CD-ROM, version 2*. Presses Universitaires de Louvain, Louvain-la-Neuve, Belgium.

Shin’ichiro Ishikawa. 2011. A New Horizon in Learner Corpus Studies: The Aim of the ICNALE Projects. In G. Weir, S. Ishikawa, and K. Poonpon, editors, *Cor-*

pora and Language Technologies in Teaching, Learning and Research. University of Strathclyde Publishing.

Scott Jarvis and Scott Crossley, editors. 2012. *Approaching Language Transfer Through Text Classification: Explorations in the Detection-based Approach*, volume 64. Multilingual Matters Limited, Bristol, UK.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 624–628, Chicago, IL. ACM.

Michael Swan and Bernard Smith, editors. 2001. *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, 2 edition.

Benjamin 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 (Volume 2: Short Papers)*, pages 193–197, Jeju Island, Korea, July. Association for Computational Linguistics.

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 COLING 2012*, pages 2585–2602, Mumbai, India, December. The COLING 2012 Organizing Committee.

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*, pages 1600–1610, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Sze-Meng Jojo Wong, Mark Dras, and Mark Johnson. 2011. Topic Modeling for Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 115–124, Canberra, Australia, December.

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*, pages 699–709, Jeju Island, Korea, July. Association for Computational Linguistics.

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: Human Language Technologies*, pages 180–189, Portland, Oregon, USA, June. Association for Computational Linguistics.