Native Language Identification Using Large, Longitudinal Data

Xiao Jiang[1], Yufan Guo[1], Jeroen Geertzen[2], Dora Alexopoulou[2], Lin Sun[1], Anna Korhonen[1]
[1] Computer Laboratory, University of Cambridge, UK
[2] Department of Theoretical and Applied Linguistics, University of Cambridge, UK
xj229@cam.ac.uk, yg244@cam.ac.uk, jg532@cam.ac.uk, ta259@cam.ac.uk, linsun84@gmail.com, alk23@cam.ac.uk

Abstract

Native Language Identification (NLI) is a task aimed at determining the native language (L1) of learners of second language (L2) on the basis of their written texts. To date, research on NLI has focused on relatively small corpora. We apply NLI to the recently released EFCamDat corpus which is not only multiple times larger than previous L2 corpora but also provides longitudinal data at several proficiency levels. Our investigation using accurate machine learning with a wide range of linguistic features reveals interesting patterns in the longitudinal data which are useful for both further development of NLI and its application to research on L2 acquisition.
Keywords: Native Language Identification, EFCamDat, L2 Acquisition

1. Introduction

Native Language Identification (NLI) is a task aimed at determining the native language (L1) of learners of second language (L2) on the basis of their written texts. NLI is important for Natural Language Processing (NLP) applications and can also offer important input to research on L2 acquisition. In particular, it can shed light on whether L1 background influences L2 learning and whether there is a significant difference between the writings of L2 learners with different L1 backgrounds and at different proficiency levels.

Since the first work of Koppel et al. (2005) on NLI, researchers have mainly treated NLI as a supervised text classification task. A variety of machine learning methods (most notably Support Vector Machines) have been explored with different linguistic features, including e.g. function words (Koppel et al., 2005), character n-grams (Tsur and Rappoport, 2007; Ahn, 2011), part-of-speech (POS) tags (Bykh and Meurers, 2012), syntactic structures (Wong and Dras, 2011), error-based features (Kochmar, 2011) and style features (e.g. sentence length) (Bergsma et al., 2012). Results have been promising, with the accuracy ranging between 72.0-94.6% depending on the task in hand.

However, the majority of NLI studies (nearly all of those focused on English as L2) have employed the International Corpus of Learner English (ICLE) (Granger, 2003) as input data. This corpus is relatively small in size (3.7M words) and restricted in terms of topics and proficiency levels covered, including data from advanced students only. Other NLI copora in use include the Cambridge Learner Corpus that contains over 200K exam scripts from students taking Cambridge English exams (Kochmar, 2011), the English Wikipedia dataset that consists of 2.4 million Wikipedia comments (Al-Rfou, 2012), the subset (years 1999-2009) of the ACL Anthology Network (Bergsma et al., 2012), along with the recently-released TOEFL11 corpus of 12,100 TOEFL iBT essays (Blanchard et al., 2013).

In this paper we employ, for the first time, the recently released EF-Cambridge Open English Learner Database (EFCamDat) corpus (Geertzen et al., 2013) for NLI. This new corpus is multiple times larger than previous L2 corpora and provides longitudinal data at several proficiency levels. It contains over 30M words of written assignments covering as many as 128 topic areas, produced by learners at 16 different proficiency levels. As the data come from a live educational context capturing writings of large numbers of students from diverse backgrounds, they provide much richer resource for the development and evaluation of NLI as well as for linguistic studies.

We explore the potential and challenges of NLI when applied to this new longitudinal data. We report experiments where we first extract a rich set of linguistic features from this corpus (including word and character n-grams, POS n-grams, production rules, and grammatical relations) using state-of-the-art NLP. We then classify the features using Support Vector Machines (SVM) a widely-used classifier which has yielded high performance in previous research on NLI. We finally go on to conduct a thorough quantitative and qualitative evaluation of the performance of different features, comparing, for the first time, performance across different proficiency levels. We observe patterns interesting for both further development of NLI and its application to research on L2 acquisition. We show that the top performing features differ across proficiency levels. We take some of these patterns forward for linguistic analysis and show the relevance of our analysis for research on L2 acquisition.

2. Data

EFCamDat was developed in the University of Cambridge, in collaboration with Education First (EF) – a world-leading company in international education. The corpus contains essays submitted to EF Englishtown[1] – an online English school offering E-learning for users at any level of proficiency. Table 1 shows the current number of documents, words, learners, nationalities and proficiency levels covered by EFCamDat (as of October 2013).

Since we wanted to investigate different proficiency levels and not all levels had sufficient data for adequate NLI performance at the time of this experiment, we merged proficiency levels into 4 groups as to avoid data sparsity, as shown in Table 2.

We focused on three major nationalities – Chinese, Brazil-

[1] http://www.englishtown.com/

	Count
Documents	423,373
Words	30,763,521
Learners	76,002
Nationalities	175
Proficiency Levels	16

Table 1: The statistics of EFCamDat.

ian[2] and Russian – which jointly cover 78% of the corpus and yield a reasonably large training and test sets for NLI. We excluded some essays to ensure that each group contains approximately the same amount of data.

Group	Documents	Words
Lvl 1-3	44,362	1,910,674
Lvl 4-7	50,593	4,223,560
Lvl 8-11	15,095	1,726,093
Lvl 12-16	2,459	359,563

Table 2: A subset of EFCamDat used in this work

3. Methods

3.1. Features

We investigated a variety of lexical and syntactic features used in previous NLI works:

Word n-gram Word n-gram is a widely used feature type in many text classification tasks including the recent NLI shared task (Tetreault et al., 2013). We experimented with word n-grams of different orders (n = 1 to 4) under different settings, such as whether to convert all letters to lower case, perform lemmatization, remove stop words and punctuation, filter out low frequency n-grams, and so on.

Character n-gram Character n-grams can capture spelling mistakes or preferences and have proved useful for NLI in (Ahn, 2011). We experimented with character n-grams in a similar way as with word n-grams.

POS n-gram POS n-grams may capture non-target distributional patterns, that can be characteristic due to a specific L1 background. For instance, a 3rd person singular noun followed by a verb base form "John love" would indicate absence of subject-verb agreement. Therefore, POS n-grams may serve as a good feature for NLI. We used the Penn Treebank POS tag set (Mitchell Marcus, 2012) in our work, and experimented with POS n-grams of different orders (n = 2 to 5).

Production rules In a formal grammar, a production rule (PR) is a rewrite rule that specifies a symbol substitution for generating new symbol sequences, e.g. *S→NP+VP*. As PR can be a good indicator of the use of grammar, we applied it to NLI in view of its high performance reported in (Wong and Dras, 2011). In addition to standard PR, we also tested its lexicalized version, with the corresponding words attached to each symbol, e.g. *S→NP_I+VP_went*.

Dependency features Dependency (or grammatical) relations are functional relationships between constituents in a clause, such as *ncsubj* for non-clausal subject relations, *dobj* for direct object relations, and so on. Dependency features provide a good representation of the syntactic structure of a sentence and are thus potentially useful for distinguishing between the writing styles of different L2 learners.

3.2. Experiment Setup

We used the Stanford parser (Klein and Manning, 2003) for feature extraction. Following Bykh and Meurers (2012), we used the LIBLINEAR SVM classifier (Fan et al., 2008) for ML-based NLI. To avoid selection bias, we performed 4-fold cross validation and reported the average accuracy at levels 1-3, 4-7, 8-11, 12-16, respectively.

4. Results

We investigated the performance of each individual feature type, as well as how they can complement each other.

4.1. Individual Features

Table 3 shows the accuracy of our NLI system when using each individual feature type alone. Here we report the results for the best configuration of the features only. For lexical features (word and character n-grams), we found that normalization or filtering resulted in a slight decrease in accuracy, and that the best configuration was to use the original form of all words or characters. The optimal N for word/character/POS n-gram features varies across different proficiency levels. In general, a larger N is preferred as the proficiency level goes up, which makes sense given that longer and more complex expressions are featured in the written work of advanced learners. For PR and Dependency features, their lexicalized version tends to perform better than their original form in most cases.

	1-3	4-7	8-11	12-16
Word	81.16%	79.17%	77.57%	63.12%
Char	81.19%	81.60%	79.30%	66.89%
POS	57.32%	62.37%	62.81%	56.29%
PR	51.88%	58.95%	60.32%	53.71%
Depd	45.16%	51.50%	55.06%	49.62%

Table 3: Performance of each individual feature type

As shown in Table 3, lexical features (word and character n-grams) significantly outperform syntactic ones (POS, production rules and dependencies) by up to 36%. For lexical features, the impact of L1 is more significant for beginners than for advanced learners. For syntactic features, the impact of L1 is clearer at medium than at low or high proficiency levels. This is probably because at the beginner levels, students are exposed to simple syntactic constructions which are relatively easy to learn, while by the time they get to the advanced levels, most students have a good grasp of grammar and their L1 background has less impact.

[2]The native language of Brazilians is Portuguese.

POS is the most telling of the three syntactic features probably because it can tap into morpho-syntax and also lexically driven patterns, whereas PR and Depd are too abstract and at that level the syntax of many languages looks rather similar.

4.2. Combination of features

We also conducted experiments to investigate how different features complement each other. Table 4 shows the results when using all but one of the features, from which we can see that lexical features contribute to overall performance in almost all cases, with the only exception for character n-gram for levels 1-3. When combined with other features, syntactic features such as PR and dependency play a less important role in NLI when students enter into more advanced levels, whereas POS n-gram becomes more and more indispensable for NLI as the proficiency level goes up.

	1-3	**4-7**	**8-11**	**12-16**
All	82.09%	82.54%	80.84%	69.50%
w/o Word	-0.66	-0.45	-0.37	-0.53
w/o Char	+0.68	-2.09	-2.11	-2.57
w/o POS	+0.73	+0.43	-0.42	-1.20
w/o PR	-0.18	-0.24	+0.48	+1.38
w/o Depd	+0.11	+0.24	+0.27	+0.40

Table 4: Accuracy gain (+%) or loss (-%) of leave-one-out experiments.

4.3. Qualitative Analysis

We selected up to 100 best-performing features for each feature type using the Information Gain (Yang and Pedersen, 1997) criteria. Table 5 shows the top 5 features for word unigrams, character 6-grams, POS bigrams, production rules, and dependencies, respectively. The top features with other configurations (e.g. n-grams of different orders) have similar trends and are not shown here.

As shown in Table 5, the most indicative features vary a lot from one proficiency level to another. Take word n-grams for example: the best-performing features for beginners are country names that express one's L1 background explicitly. This is probably because most L2 learners at the beginner levels are told to write a self introduction that includes sentences such as "I am from Russia" or "I live in São Paulo'. As they enter into more advanced levels, students become more experienced in using function words e.g. *which* and *that* to write more complex sentences, and students with different L1 backgrounds may have different preferences of words. As a second example, the best-performing PR features seem to shift from the phrase level (e.g. *NP→NNP+FW+NNP*) to the sentence level (e.g. *S→PP+NP+VP*) as the learners become more proficient. These results suggest that features corresponding to more complex structure of sentences tend to be more informative for NLI in advanced learners.

We also had an experienced linguist to perform a qualitative analysis of these representative features. Some of our findings are summarized below:

	Word n-grams (n = 1)
Lvl 1-3	*russia*; *brazil*; *china*; *moscow*; *paulo*
Lvl 12-16	*which*; *brazil*; *that*; *it*; *suitable*
	Char n-grams (n = 6)
Lvl 1-3	*Russia*; *m_Russ*; *China_*; *om_Bra*; *m_Bras*
Lvl 12-16	*which*; *brazil*; *becaus*; *As_for*; *_suita*
	POS n-grams (n = 2)
Lvl 1-3	*FW NNP*; *NNP FW*; *NNP NNP*; *MD VB*; *PRP MD*
Lvl 12-16	*COMMA PRP*; *COMMA IN*; *COMMA CC*; *NNS DOT*; *NN PRP*
	Production Rules
Lvl 1-3	*NP→NNP+FW+NNP*; *VP→MD+VP*; *VP→MD+RB+VP*; *PP→IN+NP*; *ADJP→NP+JJ*
Lvl 12-16	*S→PP+NP+VP*; *S→CC+NP+VP*; *S→S+CC+S*; *S→SBAR+NP+VP*; *S→ADVP+NP+VP*
	Dependencies
Lvl 1-3	*neg*; *npadvmod*; *ccomp*; *det*; *prep_opposite*
Lvl 12-16	*prepc_as_for*; *prep_about*; *prep_of*; *prep_from*; *preconj*

Table 5: The top 5 features for different feature types and proficiency levels. Please refer to the Penn Treebank POS tag set (Mitchell Marcus, 2012) and the Stanford parser (De Marneffe and Manning, 2008) for meanings of POS tags, production rules and typed dependencies. Underscores "_" in character 6-grams represent a white space between words. COMMA and DOT refer to punctuations comma and dot.

Lexical preferences Certain punctuation marks and phrases are used more frequently by English learners from one country than another. For instance, Chinese students do not use dashes as frequently as Russians or Brazilians do. As another example, phrases such as *"as for me"* and *"to my mind"* are featured in the essays of Russian students, and phrases such as *"try my best"* and *"what's more"* are commonly used by Chinese students, perhaps due to the frequent use of the same expression in Russian and Chinese languages.

Clause-initial prepositional phrase An interesting feature that is typically useful for distinguishing between Chinese, Russian and Brazilian students is the clause-initial prepositional phrase (PP), such as "*In the afternoon* he goes shopping at 3 o ' clock".

For instance, at levels 1-4, Chinese students tend to put time references at the beginning of a clause to emphasize its tense, e.g. "*On Sunday*, he goes to the park and meets friends, and at half past eleven he plays tennis with his friends." This may be because Chinese learners rely on structural means to communicate temporality in their L1.

As students enter into advanced levels, they become more experienced in using different verb forms (e.g. affixes) as tense markers, and the use of clause-initial PP for time reference is significantly reduced.

Russians also use clause-initial PPs for temporal reference. But their phrases are more complex, including often two points of temporal reference, e.g. "*On Saturday at eleven thirty*". In addition, they use clause-initial PPs to indicate not just time, but also location and manner e.g. "*With great pleasure* we inform our clients that ...". This is strongly correlated with their L1 background as clause-initial PPs are frequently used in the Russian language (King, 1995).

5. Conclusion

We have developed a method for NLI which employs accurate machine learning (SVM) with a wide range of linguistic features (ranging from character features to syntactic dependencies) and have applied this method to the newly developed, large EFCamDat corpus which, unlike previous learner corpora, provides longitudinal data at multiple proficiency levels. We have performed, for the first time, an experiment where we compare the performance at different proficiency levels. We report high overall accuracy of around 80% at low and medium proficiency levels, and 70% at advanced levels. Our quantitative and a qualitative analysis of different features reveals that the top performing features differ from one proficiency level to another. Our linguistic analysis shows that our results can be of interest to research on L2 acquisition.

In the future, we plan to investigate NLI at finer-grained levels of proficiency and to integrate a wider range of nationalities, exploring strategies to deal with data sparsity. We also plan to develop new NLI methodology suitable for the analysis of large, longitudinal data, based on the insights gained in our experiments. Finally, we plan to conduct further linguistic evaluation of the data.

Acknowledgments

We thank EF Education First for providing the data and sponsoring the development of EFCAMDAT, the Isaac Newton Trust (Cambridge) for a grant supporting the development of EFCAMDAT, and finally the Royal Society, UK.

6. References

Ahn, C. S. (2011). *Automatically detecting authors' native language*. Ph.D. thesis, Monterey, California. Naval Postgraduate School.

Al-Rfou, R. (2012). Detecting english writing styles for non-native speakers. *arXiv preprint arXiv:1211.0498*.

Bergsma, S., Post, M., and Yarowsky, D. (2012). Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). Toefl11: A corpus of non-native english. *Educational Testing Service*.

Bykh, S. and Meurers, D. (2012). Native language identification using recurring n-grams–investigating abstraction and domain dependence.

De Marneffe, M.-C. and Manning, C. D. (2008). Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Geertzen, J., Alexopoulou, T., Baker, R., Hendriks, H., Jiang, S., Korhonen, A., and First, E. E. (2013). The ef cambridge open language database (efcamdat) user manual part i: written production.

Granger, S. (2003). The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 37(3):538–546.

King, T. H. (1995). *Configuring Topic and Focus in Russian*. CSLI Publications.

Klein, D. and Manning, C. D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Kochmar, E. (2011). *Identification of a writer's native language by error analysis*. Ph.D. thesis, Master's thesis, University of Cambridge.

Koppel, M., Schler, J., and Zigdon, K. (2005). Automatically determining an anonymous authors native language. In *Intelligence and Security Informatics*, pages 209–217. Springer.

Mitchell Marcus, Ann Taylor, R. M. (2012). Alphabetical list of part-of-speech tags used in the penn treebank project.

Tetreault, J., Blanchard, D., and Cahill, A. (2013). A report on the first native language identification shared task. *NAACL/HLT 2013*, page 48.

Tsur, O. and Rappoport, A. (2007). Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16. Association for Computational Linguistics.

Wong, S.-M. J. and Dras, M. (2011). Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610. Association for Computational Linguistics.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 412–420. MORGAN KAUFMANN PUBLISHERS, INC.