Large-scale Native Language Identification with Cross-Corpus Evaluation

Shervin Malmasi
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

Abstract

We present a large-scale Native Language Identification (NLI) experiment on new data, with a focus on cross-corpus evaluation to identify corpus- and genre-independent language transfer features. We test a new corpus and show it is comparable to other NLI corpora and suitable for this task. Cross-corpus evaluation on two large corpora achieves good accuracy and evidences the existence of reliable language transfer features, but lower performance also suggests that NLI models are not completely portable across corpora. Finally, we present a brief case study of features distinguishing Japanese learners' English writing, demonstrating the presence of cross-corpus and cross-genre language transfer features that are highly applicable to SLA and ESL research.

1 Introduction

Native Language Identification, the task of determining the native language (L1) of an author based on a second language (L2) text, has received much attention recently. Much of this is motivated by Second Language Acquisition (SLA) as NLI, often accomplished via machine learning methods, can be used to study language transfer effects.

Most NLI research hitherto has focused on identifying linguistic phenomena that can capture transfer effects, with little effort towards interpreting discriminant features. Some researchers have now shifted their focus to developing data-driven methods for the automatic extraction and ranking of linguistic features that distinguish specific L1s (Swanson and Charniak, 2014).

Such methods could be used not only to confirm existing SLA hypotheses, but also to create new ones. This hypothesis formulation is an inherently difficult problem requiring copious amounts of data. Contrary to this requirement, researchers have long noted the paucity of suitable corpora[1] for this task (Brooke and Hirst, 2011). This is one of the research issues addressed by this work.

Furthermore, deriving SLA hypotheses from a single corpus may not be entirely useful for SLA research. Many variables like genre and topic are constant within a corpus, restricting the validity of such cross-validation studies to those dimensions.

An alternative, potentially more helpful approach, is to identify transfer features that reliably distinguish an L1 across multiple corpora of differing genres and domains. A cross-corpus methodology may be a more promising avenue to finding features that generalize to diverse text sources, but requires additional large corpora. It is also a more realistic approach, and one we pursue in this work.

Accordingly, the aims of the present work are to: (1) test a large new corpus suitable for NLI, (2) perform within-corpus evaluation with a comparative analysis against equivalent corpora, (3) perform cross-corpus evaluation to determine the efficiency of corpus independent features and (4) analyze the features' utility for SLA & ESL research.

2 Background and Motivation

NLI work has been growing in recent years, using a wide range of syntactic and more recently, lexical features to distinguish the L1. A detailed review of NLI methods is omitted here for reasons of space, but a thorough exposition is presented in the report from the very first NLI Shared Task that was held in 2013 (Tetreault et al., 2013).

Most English NLI work has been done using two corpora. The *International Corpus of Learner En-*

[1] An ideal NLI corpus should have multiple L1s, be balanced by topic, proficiency, texts per L1 and be large in size.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 1403–1409,
Denver, Colorado, May 31 – June 5, 2015. ©2015 Association for Computational Linguistics

glish (Granger et al., 2009) was widely used until recently, despite its shortcomings[2] being widely noted (Brooke and Hirst, 2012a). More recently, TOEFL11, the first corpus designed for NLI was released (Blanchard et al., 2013). While it is the largest NLI dataset available, it only contains argumentative essays, limiting analyses to this genre.

Research has also expanded to use non-English learner corpora (Malmasi and Dras, 2014a; Malmasi and Dras, 2014c). Recently, Malmasi and Dras (2014b) introduced the Chinese Learner Corpus for NLI and their results indicate that feature performance may be similar across corpora and even L1-L2 pairs. This is a claim that we will test here.

NLI is now also moving towards using features to generate SLA hypotheses. Swanson and Charniak (2014) approach this by using both L1 and L2 data to identify features exhibiting non-uniform usage in both datasets, creating lists of candidate transfer features. Malmasi and Dras (2014d) propose a different method, using linear SVM weights to extract lists of overused and underused linguistic features for each L1 group.

Cross-corpus studies have been conducted for various data-driven NLP tasks, including parsing (Gildea, 2001), WSD (Escudero et al., 2000) and NER (Nothman et al., 2009). While most such experiments show a drop in performance, the effect varies widely across tasks, making it hard to predict the expected drop for NLI. We aim to address this question using large training and testing data.

3 EFCamDat: A new corpus for NLI

The EF Cambridge Open Language Database (EFCAMDAT) is an English L2 corpus that was released recently (Geertzen et al., 2013). It is composed of texts submitted to *Englishtown*, an online school used by thousands of learners daily.

This corpus is notable for its size, containing some 550k texts from numerous nationalities, making it an ideal candidate for NLI research. While the TOEFL11 is made of argumentative essays, EFCAMDAT has a much wider range of genres including writing emails, descriptions, letters, reviews, instructions and more.

In this work we present an application of NLI to this new data. As some of the texts can be short, we use the methodology of Brooke and Hirst (2011) to concatenate and create texts with at least 300 tokens, much like the TOEFL11.

Common	Arabic, Chinese, French, German, Italian, Japanese, Korean, Spanish, Turkish
EFCAMDAT	Portuguese, Russian
TOEFL11	Hindi, Telugu

Table 1: The 11 L1 classes extracted from the EFCAMDAT corpus, compared to the TOEFL11 corpus. The first 9 classes are common between both.

From the data we choose 850 texts from each of the top 11 nationalities. This subset of EFCAMDAT thus consists of 9,350 documents totalling approximately 3.2m tokens. This is an average of 337 tokens per text, close to the 348 tokens per text in TOEFL11.

This also provides us with the same number of classes as the TOEFL11, as shown in Table 1, facilitating direct performance comparisons. The table also indicates the 9 classes common to both corpora. This subset of common classes enables us to perform large-scale cross-corpus validation experiments that have not been possible until now.

4 Methodology

We use the standard NLI classification approach. A linear Support Vector Machine is used for classification and feature vectors are created using relative frequency values. We also combine features with a mean probability ensemble classifier (Polikar, 2006, §4.2) using the probabilities assigned to each class. We compare results with a random baseline and the oracle baseline used by Malmasi et al. (2015). The oracle correctly classifies a text if any ensemble member correctly predicts its label and defines an upper-bound for classification accuracy. We avoid using lexical features as EFCAMDAT is not topic balanced. We extract the following topic-independent feature types:

Function words are topic-independent grammatical words such as prepositions which indicate the relations between other words. They are known to be useful for NLI. Frequencies of 400 English function words[3] are extracted as features. We also apply function word bigrams as described in Malmasi et al. (2013).

Context-free Grammar Production Rules are extracted after parsing each sentence. Each rule is a classification feature (Wong and Dras, 2011) and captures global syntactic patterns.

[2] The issues exist as the corpus was not designed for NLI.

[3] Like previous work, this also includes stop words, which we sourced from the Onix Text Retrieval Toolkit: http://www.lextek.com/manuals/onix/stopwords1.html

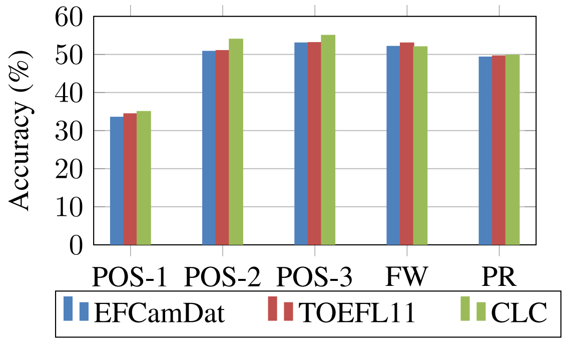

Figure 1: Comparing EFCAMDAT feature performance with the TOEFL11 and Chinese Learner Corpus (CLC). POS-1/2/3: POS uni/bi/trigrams, FW: Function Words, PR: CFG Productions

Part-of-Speech (POS) n-grams of size 1–3 are extracted as features. They capture preferences for word classes and their localized ordering patterns.

5 Within-Corpus Evaluation

Our first experiment applies 10-fold cross-validation within the corpus to assess feature efficacy. The results are shown in the first column of Table 2.

All features perform substantially higher than the 9% baseline. POS trigrams are the best single feature (53%), suggesting there exist significant interclass syntactic differences. Next, we also combined all features using a classifier ensemble, which has been shown to be helpful for NLI (Tetreault et al., 2012). This yields the best accuracy of 65% against an upper-bound of 87% set by the oracle.

We also compare these results to those from the TOEFL11 and Chinese Learner Corpus (CLC). As shown in Figure 1, we find that feature performance is nearly identical across corpora. Consistent with the results in Malmasi and Dras (2014b), this seems to suggest an invariant degree of transfer across different learners and L1-L2 pairs.

Figure 2 shows the confusion matrix. German is the most correctly classified L1, while the highest confusion is between Japanese–Korean, followed by Spanish–Portuguese and French–Italian. This is not surprising given their syntactic similarity as well as being typologically related in case of the latter two.

6 Large-scale Cross-Corpus Evaluation

Our second experiment tests the cross-corpus efficacy of the features by training on EFCAMDAT and testing on TOEFL11,[4] and *vice versa*. As the corpus texts are from different genres, this approach enables us to test the cross-corpus and cross-genre generalizability of our features.

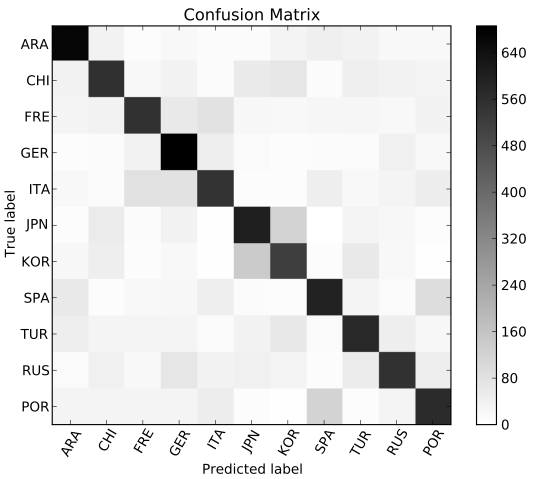

Figure 2: EFCAMDAT 11-class confusion matrix.

Arabic	German	Japanese
Saudi	Germany	Japan
Arabia	Berlin	Tokyo
Arabic	Hamburg	Osaka
Mohammed	Frankfurt	Nagoya
Ali	Munich	Yen

Table 3: Selected items from the top 15 most discriminative words for Arabic/German/Japanese.

Results are shown in the second and third column of Table 2. While lower than the cross-validation results which were on 11 classes *vs* 9 here, the results are far greater than the baseline. The accuracy for training on EFCAMDAT and testing on TOEFL11 is higher (33.45%) than the other way around (28.42%), even though TOEFL11 is the larger corpus. This is possibly because EFCAMDAT has numerous genres while TOEFL11 does not. The cross-corpus oracle is also over 20% lower, despite an increase in the random baseline, showing some features are not portable across corpora. Training on TOEFL11 yields a lower oracle.

Although a performance drop was expected due to the big genre differences, results suggest the presence of some corpus-independent features that capture cross-linguistic influence. However, they also suggest that a large portion of the features helpful for NLI are genre-dependent.

Previously, word n-grams have been applied in small-scale cross-corpus studies and found to be the best feature (Brooke and Hirst, 2012b). Word n-grams have been previously used in NLI and are believed to capture lexical transfer effects which have been previously noted by researchers and linguists

[4] The 9 common classes discussed in §3 are used.

Classification Feature	EFCAMDAT 10-fold CV	Train EFCAMDAT Test TOEFL11	Train TOEFL11 Test EFCAMDAT
Random Baseline	9.09	11.11	11.11
Oracle Baseline	86.84	64.92	62.43
Function Word unigrams	52.01	27.14	21.77
Function Word bigrams	47.92	29.21	22.63
Production Rules	49.12	30.73	23.91
Part-of-Speech unigrams	33.21	23.42	16.71
Part-of-Speech bigrams	50.43	31.02	23.09
Part-of-Speech trigrams	53.05	32.38	25.55
Ensemble (All features)	**64.95**	**33.45**	**28.42**
Word unigrams	–	41.82	42.48

Table 2: Classification accuracy (%) for our within- and cross-corpus experiments.

(Odlin, 1989). The effects are mediated not only by cognates and word form similarities, but also semantics and meanings. Other NLI studies have also provided empirical evidence for this hypothesis (Malmasi and Cahill, 2015).

However, issues stemming from topic bias[5] have also limited their use in NLI (Brooke and Hirst, 2012a), although use could be justified in cross-corpus scenarios due to the lower risk of topic-bias across corpora. We applied word unigrams to our cross-corpus experiment, achieving an accuracy of 41.8% for training on the EFCAMDAT and testing on TOEFL11 and 42.5% for the reverse setting. These are the best results in this setup.

To check for any topic-bias effects, we inspected the most discriminative features for each L1 class using the method proposed by Malmasi and Dras (2014d). This analysis revealed that the top features were mostly cultural and geographic references related to the author's country. Table 3 contains words selected from the top 15 most discriminative features found in the cross-corpus experiment for three L1s. We observe that most of these are toponyms or culture-specific terms such as names and currencies. These results reveal another potential issue with using lexical features. Although this isn't topic-bias, the features do not represent genuine linguistic differences or lexical transfer effects between L1 groups. In practical scenarios, this could also make NLI systems vulnerable to content-based manipulation. The exclusion of proper nouns is one way to combat this.

7 A Case Study of Japanese Learners

To demonstrate the utility of this cross-corpus approach we present a brief case study of features that characterize English writings of Japanese learners. We extracted the most discriminative cross-corpus features of Japanese learner texts using the method of Malmasi and Dras (2014d).

Table 4 contains the top production rule features. The first rule shows a preference for having a subordinate clause before the main clause. The next two rules show that Japanese learners tend to begin their sentences with adverbs and conjunctions. This preference for placing information at the start of sentences is most likely rooted in the fact that Japanese is an SOV head-final language[6] where dependent clauses generally precede the main clause and relative clauses precede the noun they modify. The influence of this head-direction parameter on English acquisition has been previously investigated (Flynn, 1989).

In contrast, it is quite common for the main clause to precede the subordinate clause in English. Other research has also noted that Japanese speakers have a "long before short" preference[7] (Yamashita and Chang, 2001). This is also evidence by another highly discriminative rule for this L1: S → S , CC S .

Japanese writers also seem more likely to split longer arguments into multiple shorter sentences, as suggested by our third production rule. It has also been noted that Japanese and Korean sentences in the TOEFL11 have the shortest mean length (Cimino et al., 2013, p. 211).

Turning to POS trigrams, the POS tag sequence VBZ JJ NN is strongly linked to Japanese learn-

[5] Due to correlations between text topics and L1 classes.

[6] Contrasted with English which is SVO.

[7] This refers to how conjuncts are ordered: short-before-long in English, long-before-short in Japanese. Our findings suggest that Japanese writers transfer this internal order-preference into their L2 English writing.

Production Rule	Example Sentence
S → SBAR , NP VP .	If you have spare time, you'll think of shopping.
S → ADVP , NP VP .	Therefore, the online studying system is really convenient for me.
S → CC NP VP .	But I'm not good at English. / But it wasn't comfortable and cosy.

Table 4: The top 3 cross-corpus production rule features for Japanese L1 with example lexicalizations.

Overuse	Underuse
however	perhaps
though	somebody
cannot	everything
therefore	behind
such	upon
into	between

Table 5: English function words overused and underused by Japanese learners in their writing.

ers. It represents a third person verb, such as *is* or *has* followed by an adjective and a noun. A brief analysis reveals that this is commonly observed in Japanese learner texts because the sequence is missing a determiner before the noun phrase.[8] This likely stems from the fact that Japanese learners have difficulty with English articles, often failing to use them (Butler, 2002; Thomas, 1989). Its prominence in the ranked list shows that it is a common issue across distinct learners and genres.

The top overused and underused function words are listed in Table 5. The words *however* and *therefore* are highly relevant; Japanese writers often use these to start sentences, possibly due to the above-mentioned production rules. The word *into* is also predictive and seems to be used in places where *in* is more appropriate. This may be due to the Japanese words for *in*, *to* and *into* being similar.[9] In the underuse list, *perhaps* is never used by Japanese learners. Other words here are low-frequency in Japanese L1 texts in both corpora.

8 Discussion

In this work we presented the first application of one of the largest and newest publicly available learner corpora to NLI. Cross-validation experiments mirrored the performance of other corpora and demonstrated its utility for the task. We believe this will motivate future work by equipping researchers with a large-scale corpus that is highly suitable for NLI.

Next, results from the largest cross-corpus NLI evaluation to date were presented, providing strong evidence for the presence of transfer features that generalize across learners, corpora, topics and genres. However, the fact that the cross-corpus accuracy is lower than within-corpus cross-validation highlights that a large portion of the features are highly corpus-specific. This suggests that NLI models are not entirely portable across corpora. Practical applications of NLI to forensic linguistics or SLA must be robust to input from numerous sources and their associated variations, and this finding highlights the need for a cross-corpus approach.

To demonstrate how this methodology could be used for SLA, an examination of the cross-corpus features effective in classifying texts of Japanese learners was conducted. Through feature analysis, we were able to link these patterns of syntactic productions, article use and lexical choices to L1-based SLA hypotheses.

Our output lists hundreds of features, not included or examined here due to space limitations, whose analysis would allow SLA researchers to explore and generate new hypotheses, specially by combining multiple syntactic feature types.

A shortcoming here is that we did not balance texts by proficiency to match the TOEFL11. We expect that a more even sampling of proficiency or using proficiency-segregated models will yield higher accuracy and features more representative of students at each proficiency level.

Directions for future work are manifold. The next phase of this research will focus on developing tools to derive and browse ranked lists of the most discriminative cross-corpus features, which will then be used to formulate SLA hypotheses. Subject to availability of data, this could be expanded to a multiple cross-corpus methodology, using three or more corpora. Its application to other languages besides English is also of interest.

NLI is a young but rapidly growing field of research and this study is but a first step in shifting efforts towards a more interpretive approach to the task. We hope that the new dataset and directions presented here will galvanize future work.

[8] Example lexicalizations from EFCAMDAT include "She *wears black top*" and "This area *is famous park*."

[9] All use the particle *ni*, see Takenobu et al. (2005)

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 *Conference of Learner Corpus Research*, University of Louvain, Belgium.

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

Julian Brooke and Graeme Hirst. 2012b. Robust, Lexicalized Native Language Identification. In *Proc. Internat. Conf. on Computat. Linguistics (COLING)*.

Yuko Goto Butler. 2002. Second language learners' theories on the use of english articles. *Studies in second language acquisition*, 24(03):451–480.

Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2013. Linguistic profiling based on general–purpose features and native language identification. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–215, Atlanta, Georgia, June. Association for Computational Linguistics.

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 172–180. Association for Computational Linguistics.

Suzanne Flynn. 1989. The role of the head-initial/head-final parameter in the acquisition of English relative clauses by adult Spanish and Japanese speakers. *Linguistic perspectives on second language acquisition*, pages 89–108.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat).

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 167–202.

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

Shervin Malmasi and Aoife Cahill. 2015. Measuring Feature Diversity in Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Denver, Colorado, June. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014a. Arabic Native Language Identification. In *Proceedings of the Arabic Natural Language Processing Workshop (EMNLP 2014)*, pages 180–186, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014b. Chinese Native Language Identification. pages 95–99, Gothenburg, Sweden, April. Association for Computational Linguistics.

Shervin Malmasi and Mark Dras. 2014c. Finnish Native Language Identification. In *Proceedings of the Australasian Language Technology Workshop (ALTA)*, pages 139–144, Melbourne, Australia.

Shervin Malmasi and Mark Dras. 2014d. Language Transfer Hypotheses with Linear SVM Weights. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1385–1390, Doha, Qatar, October. Association for Computational Linguistics.

Shervin Malmasi, Sze-Meng Jojo Wong, and Mark Dras. 2013. NLI Shared Task 2013: MQ Submission. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–133, Atlanta, Georgia, June. Association for Computational Linguistics.

Shervin Malmasi, Joel Tetreault, and Mark Dras. 2015. Oracle and Human Baselines for Native Language Identification. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, Denver, Colorado, June. Association for Computational Linguistics.

Joel Nothman, Tara Murphy, and James R Curran. 2009. Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 612–620. Association for Computational Linguistics.

Terence Odlin. 1989. *Language Transfer: Cross-linguistic Influence in Language Learning*. Cambridge University Press, Cambridge, UK.

Robi Polikar. 2006. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45.

Ben Swanson and Eugene Charniak. 2014. Data driven language transfer hypotheses. *EACL 2014*, page 169.

Tokunaga Takenobu, Koyama Tomofumi, and Saito Suguru. 2005. Meaning of japanese spatial nouns. In *Proceedings of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 93–100.

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.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A Report on the First Native Language Identification Shared Task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57, Atlanta, Georgia, June. Association for Computational Linguistics.

Margaret Thomas. 1989. The acquisition of English articles by first-and second-language learners. *Applied Psycholinguistics*, 10(03):335–355.

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.

Hiroko Yamashita and Franklin Chang. 2001. “Long before short” preference in the production of a head-final language. *Cognition*, 81(2):B45–B55.