L1ML: Native language identification using TOEFL11

Yuzhe Chen, Jessica Huynh, Ryan Nicoll
Brandeis University
Waltham, Massachusetts USA
{yzchen,jhuynh,rnicoll}@brandeis.edu

Abstract

In this paper, we outline a corpus linguistics-based approach to the task of native language identification (NLI) of L2 writers of English using the TOEFL11 corpus. Previous use of the TOEFL11 corpus for NLI used structural features such as characters, word length, and n-grams for classification features (Tetreault et al., 2013). To expand upon this research, we provide a description of L1ML, a specification language for annotation of English L2 morphosyntactic errors in noun and verb argument structure and agreement. We demonstrate the utility of this mark-up on a modified gold standard version of the TOEFL11 corpus, where we have provided additional controls for language, question prompt, and score. Finally, we use a Naïve Bayes classifier to show how the addition of features from L1ML can provide improvement from bare structural features alone.

Keywords: Native language identification, annotation, English as a second language

1. Introduction

In an increasingly interconnected world, languages spoken on a global scale, such as English, have a large number of non-native (NN) speakers who demonstrate systematic linguistic errors in their English in the process of acquiring the language as an adult. Given research that these errors may be dependent on their first language, or L1 (Tetreault et al., 2013), we set out to show that it may be possible to automatically identify the L1 of these writers from the language-specific errors that they make.

The study of native language identification (NLI) has been a recently growing area of interest in the field of natural language processing. It has potential applications in a wide variety of fields, including international security, data mining/advertising, second language learning, automatic error correction, among many others.

The L1ML group seeks to capitalize on the errors of NN writers of English through additional annotation following the L1ML, a mark-up language that encodes errors in spelling, punctuation use, and in agreement among nouns, verbs, and other predicates and their corresponding determiners and prepositions. Specifically, we aim to suggest that the addition of annotation can provide robust, salient information that can improve on existing structural features.

2. Related Work

A comprehensive review of NLI literature is beyond the scope of this paper; however, we find it prudent to mention a few key academic highlights that have informed our research and steered the course of this project.

The 2013 NLI shared task used the TOEFL11 corpus (described in more detail in Section 3.1.) which we have implemented in our own form for the purposes of this assignment. Furthermore, various research teams within the task used a litany of machine learning and feature extraction methods to train their classifiers, including part-of-speech n-grams, character unigrams, and spelling and grammatical errors.

3. Experimental Setup

3.1. Corpus

We have used the TOEFL11 corpus compiled by the Educational Testing Service for our research with some modifications to control for additional variables in testing and a reduction in the universe of languages (which we will called modified TOEFL7, or M-TOEFL7). The TOEFL, or Test of English as a Foreign Language, is an entrance examination akin to the SAT that measures the academic English ability of NN English speakers who wish to enter American universities.

The original TOEFL11 corpus is comprised of 12,100 open-response unstructured written answers from 11 different languages to 8 general, non-domain specific questions, such as: *Do you agree or disagree with the following statement? Young people nowadays do not give enough time to helping their communities. Use specific reasons and examples to support your answer.* Each essay is given a score from a low 0 to a high of 5 from three raters, which is then synthesized to produce a global essay score (Blanchard et al., 2013).

This corpus has several useful characteristics that make it especially applicable for this purpose. Perhaps the most important is that it has a wide variety of NN speakers from various L1s (whose L1s are already classified and encoded in the metadata) who are performing a standardized task. Previous researchers have noted the difficulty of compiling such a corpus and the paucity of data available for NLI research that allows for cross-L1 NLI classification (Blanchard et al., 2013). Additionally, the TOEFL11 controls for language and prompt by taking a relatively even sampling per language and per prompt. However, it does not currently control for global essay score; consequently, the L1ML team set to modify the corpus to control for score and account for other variables salient to our markup goals.

For the M-TOEFL7 corpus, we have reduced the universe of languages from 11 to 7 languages: Arabic, Chinese, French, Hindi, Japanese, Telugu, and Spanish. This reduction was instituted in order to provide sufficient time for our volunteer annotators to be able to produce enough data to

train our machine learning models (though see discussion in Section 5 for why may not have been as effective as it could be along with possible solutions). These languages were also chosen for their relative difficulty in distinguishing among them for a Naïve Bayes classifier baseline across the modified version of our corpus (see Section 4.2. for relevant tables and discussion).

M-TOEFL7 also limited the scope of English proficiency relevant to our current task. In addition to the global essay scores mentioned above, the original compilers of the TOEFL11 corpus categorized each essay into 3 proficiency scores: Low (0–2.5), Medium (2.5–3.5), and High (3.5–5.0). In initial analyses of the data, the L1ML team came to two conclusions. First, the essays in the High categories were virtually indistinguishable from native speaker essays, apart from perhaps one or two small, questionable errors that the three researchers often themselves could not come to agreement on. Second, the Low essays often were short, off-task, and filled with enormous amounts of errors. Part of our annotation task is for our annotators to annotate their corrections (i.e., "repair" the errors in the sentence). "Repair" as a concept will be discussed more in-depth in Section 3.2., but suffice it to say here that it requires that the annotators have a model of what the "intended" grammatical sentence of the NN writer was. One of the difficulties the L1ML group encountered was that no standard agreement could be found among what the intended utterance was for most of the sentences in the Low category, given that many of the sentences were semantically/pragmatically bizarre or unmeaningful.

Given these observations, the L1ML group narrowed the purview of English proficiency to Medium scores, which during initial annotation rounds seemed to have a task-significant number of errors while still allowing annotators to come to agreement on intended utterances for the task. Controlling for score, question prompt, and language, our final M-TOEFL7 corpus ended up with 16 documents for each of Arabic, French, Spanish, Chinese, Japanese, and Hind and 10 documents for Telugu (this last quirk is a product of annotator difficulty—see Section 3.2. for further discussion). As above, this number of documents was chosen to accommodate our annotators? schedules to provide enough time for our annotators to produce accurate, comprehensive annotations (see Section 5. for discussion on improvements to this process).

3.2. Annotation

Three annotators served as volunteers for our project. Annotators were MA candidates in Computational Linguistics enrolled in COSI 140: Natural Language Annotation for Machine Learning. Annotators had three weeks to annotate approximately 112 documents (16 documents for each of 7 languages) after a first week of annotation where major changes were made to the annotation specification and overall annotation process. The number of annotators was reduced to two due to major illness on the part of one of one annotator.

Each annotator was provided with 2 to 3 languages per week, each separated into a folder of 16 documents delineated by the L1 of the writer. Annotators were also provided with copies of each of the 8 prompts, the annotation specification/guidelines, a copy of the .dtd file to load extent and link tags, and access to the L1ML group Dropbox and GitHub[1] repositories. Annotators were initially briefed on the initial specification during a meeting of COSI 140. However, this specification and certain key conventions changed somewhat from the first week of data processing (see discussion in Section 5. for how this may have impacted our data collection scheme/results). Annotations were performed in the Multi-document Annotation Environment (MAE)[2], which annotators could download and access from their personal computers.

Annotators were given an e-mail (Ryan?s) to direct pertinent questions to during the annotation process. Periodically, a summary of the answers to these questions would be sent out to all annotators via e-mail that summed up the answers to these questions. Answers to these questions (particularly those that clarified unclear information in the guidelines) were subsequently added to an updated version of the specification.

Our annotations specification gave detailed guidelines on how to mark various errors. Errors were spread across several domains, each with their own extent tags (i.e., tags that consumed lexical items or phrases):

- **Nouns**: nouns were encoded as extent tags [NOUN] and had attributes for issues with determiners, prepositions, and gender/number agreement. Determiners also had an extent tag [DETERMINER] with attributes to choose the correct determiner from a possible list.
- **Verbs**: Verbs were encoded as two extent tags – one for verbs [VERB] and their auxiliaries and another to mark the presence of a missing copula [MISSING-COPULA]. The VERB tag had attributes for a missing subject, problems with tense/form, and issues with prepositions.
- **Adjectives**: Predicative adjectives were encoded as extent tags [ADJ] and had attributes for prepositional issues.
- **Other issues**: three extent tags for misspellings [MISSPELLING], incorrectly used punctuation [AWKWARDPUNCTUATION] and a global catch-all for any seemingly semantically/pragmatically bizarre phrasing [USAGE ERROR]

In addition to the above tags, several link tags were used to capture agreement and relations between **two lexical items**, often in the form of syntactic argument structure or dependency.

- **Determiner/noun agreement**: Determiner/noun agreement (or problems thereof) was encoded using the *DetNounLink* link tag, which would link a determiner (if present) and its corresponding noun that it mismatched with.

[1]The GitHub repository is located at `https://github.com/brandeis-cosi140b-s16/L1ML`

[2]MAE can be downloaded from `http://keighrim.github.io/mae-annotation/`

- **Subject/verb agreement**: Subject/verb agreement (or problems thereof) was encoded using the *SVDisagreement* link tag, which would link a subject and its corresponding noun that it mismatched with.

- **Preposition/complement**: Issues with a preposition and the corresponding lexical item that selected for it were encoded using the *PrepositionLink* link tag, which would link a preposition to the word that selected for its form.

More specific information about the specification can be found within the document itself, which has been included on the L1ML group?s GitHub repository.
Underlying the specification and its choices is the idea of teasing out the *intended utterance* behind a given sentence in the text. With an NN English sentence with grammatical errors, we posit that there was an underlying idea intended by the speaker that could have been expressed using grammatical English, and that, in using our specification, there should be an algorithmic approach to categorizing what that underlying intended utterance might have been. Our specification tells annotators to rely on their intuition as native speakers in tandem with evidence provided by other grammatical features in the sentence first for resolving grammatical errors and subsequently encoding them using L1ML; however, rules of thumb are given (i.e., if it is unclear whether an error is in the determiner or the noun [or in a head from any one of its dependents] encode it in the head) to resolve potentially ambiguous utterances and improve inter-annotator agreement.

3.3. Adjudication and Gold Standard

The gold standard corpus is comprised of 106 documents (16 documents of Arabic, French, Spanish, Hindi, Chinese, and Japanese; 10 from Telugu) that was collated from 212 documents from two of our annotators. Each of the researchers in the L1ML team contributed to the adjudication process, following the specification as the guideline for deciding among the most appropriate tags. Using the adjudication protocol built into MAE, researchers chose the tags that most appropriately reflected the types of errors solicited from the L1ML specification. While more discussion on specific annotator experiences and agreement is to follow in Section 4.1., the majority of our gold standard (approximately 95%) was built off of the errors captured by one annotator, who anecdotally reported following closely to the guideline and weekly updates.
Each researcher adjudicated between two to three language sets. These sets were then collated in order to form the gold standard corpus. Tags were incorporated if they followed along the guideline and discarded if they did not, even if there was agreement among the annotators (i.e., marking capitalization as errors, even though the guideline expressly says not to do this).

4. Results

4.1. Inter-Annotator Agreement

Inter-annotator agreement was calculated using Krippendorff's alpha, which was chosen based on its utility for sparsely annotated text and based on the fact that our data set was smaller than originally expected. This calculation was performed using the built-in alpha-μ calculator present in MAE. The scores for inter-annotator agreement between our two annotators on our gold standard corpus is included in Table 1. Cross-tag and tag-level segmentation across the entire gold-standard and within languages was calculated. Given the relatively low numbers for our agreement, agreement among link tags was not calculated.
Agreement for MISSPELLING tags was higher given its preprocessing—most spelling errors were pre-tagged using a proprietary spell-checker. However, this agreement did not approach perfect scores due to the fact that annotators were required to find other errors, such as errors in homophones and lexical usage that might not be picked up by a spell checker.
Other tags, not having been preprocessed, were concomitantly much lower. Apart from issues with the annotators and the specification (both of which will be expounded upon further in Section 5.), there are several reasons why this may have been the case. First, many tags had partial overlap, rather than complete overlap, where the two annotators chose different spans of text of phrases of similar functions (i.e., one annotator chosen an entire verb and its modals and auxiliaries, while another just chose the "main" verb). Furthermore, a not-insignificant portion of tags (ADJ, AWKWARDPUNCTION, MISSINGCOPULA, and USAGEERROR) were used either highly infrequently or not at all in the gold standard corpus, leading to Krippendorff's alpha scores of 0 (or, bafflingly, even negative).

4.2. Machine Learning Baseline

Preliminary testing of the M-TOEFL7 corpus without additional markup showed strong results with a Naïve Bayes classifier alone. Indeed, a Naïve Bayes classifier has been shown to have a stronger performance and higher bias in smaller sets of data (Forman and Cohen, 2004).
We trained our baseline classifier on 106 documents controlling for language. Data was divided into train : test data in an 8 : 2 ratio. The baseline results confusion matrix is shown below. To further control for possible over-biases in the baseline classifier, identifying information within the text (such as names of places or mentions of native language) were removed.
The baseline classifier's result of 0.15 with a macro-averaged F1 precision approximating 0.12 suggest that a bag of words approach hovers around chance with our data set. This is in contrast to its performance on the larger data set, where its overall accuracy hovered around 65-70% per language.

4.3. Feature Extraction and Most Salient Features

Various combinations of error features were tested to determine the most useful added annotation data for improvement on the baseline. Using a tool to determine salient features built into the Natural Language Toolkit (NLTK), we determined that the features that provided the greatest improvement in training the Naïve Bayes classifier were counts of the tag attributes within Noun, Verb, and Mis-

	Overall	Misspelling	Missing Copula	Awkward Punc	Usage Error	Adjective	Det	Prep	Noun	Verb
All	0.4458	0.8239	0.2902	0.0240	0.0027	-0.0010	0.0790	0.0602	0.2068	0.1846
Arabic	0.3297	0.8508	-0.0001	-0.0006	-0.0087	-0.0026	0.0883	0.1696	0.2112	0.2771
French	0.4282	0.7147	-0.0001	-0.0003	-0.0042	-0.0009	0.0426	0.0188	0.1842	0.2818
Hindi	0.4062	0.7610	0.3075	-0.0007	-0.0039	-0.0016	0.0019	-0.0024	0.1412	0.1635
Japanese	0.5044	0.8619	-0.0001	0.1247	-0.0054	-0.0013	0.0797	0.0430	0.2685	0.1607
Spanish	0.6560	0.9170	0.0000	-0.0002	-0.0075	-0.0007	0.1263	-0.0022	0.1737	0.1056
Telugu	0.3665	0.6209	0.0000	-0.0006	0.3918	-0.0008	0.1292	0.1390	0.1953	0.0791
Chinese	0.5176	0.9105	0.5832	0.2499	-0.0053	-0.0007	0.0700	0.0294	0.2329	0.2038

Table 1: Inter-annotator agreement scores by tag

Confusion matrix

	ara	fra	hin	jpn	spa	tel	zho
ara	<.>	1	.	.	1	.	.
fra	.	<.>	.	.	.	.	.
hin	.	1	<.>	.	.	.	.
jpn	1	.	.	<1>	.	.	.
spa	.	.	1	.	<.>	.	.
tel	.	1	2	2	2	<2>	3
zho	2	.	.	.	.	.	<.>

Measures

ara	*Precision*: 0.000	*Recall*: 0.000
fra	*Precision*: —	*Recall*: 0.000
hin	*Precision*: 0.000	*Recall*: 0.000
jpn	*Precision*: 0.500	*Recall*: 0.333
spa	*Precision*: 0.000	*Recall*: 0.000
tel	*Precision*: 0.167	*Recall*: 1.000
zho	*Precision*: 0.000	*Recall*: 0.000

Accuracy: 0.150 *Macro-averaged* F_1: 0.127

Table 2: Baseline classifier using 106 documents, omitting annotations and names of countries and places

Confusion matrix

	ara	fra	hin	jpn	spa	tel	zho
ara	<1>	1	.	.	.	.	.
fra	.	<.>	1	.	.	.	.
hin	.	.	<1>	.	.	.	.
jpn	.	1	.	<1>	1	.	1
spa	.	.	.	.	<1>	1	.
tel	1	.	.	.	.	<.>	.
zho	1	1	1	2	1	1	<2>

Measures

ara	*Precision*: 0.300	*Recall*: 0.333
fra	*Precision*: 0.000	*Recall*: 0.000
hin	*Precision*: 1.000	*Recall*: 0.333
jpn	*Precision*: 0.250	*Recall*: 0.333
spa	*Precision*: 0.500	*Recall*: 0.333
tel	*Precision*: 0.000	*Recall*: 0.000
zho	*Precision*: 0.222	*Recall*: 0.667

Accuracy: 0.300 *Macro-averaged* F_1: 0.316

Table 3: Baseline classifier using annotated data and Noun, Verb, and Misspelling errors as features

spelling tags contributed to a significant improvement in accuracy, precision, and recall for all languages.

Other combinations, such as counts of all tags in general, presence/absence of various tags, count of attributes/links produced results with as high an accuracy.

5. Discussion and Conclusion

Initial results suggest an improvement from baseline results using counts of various types of features; however, there are several mitigations to make to this claim. First, the data set is far too small to be able to extrapolate any real meaningful conclusions from it, nor is it possible to say that these same results would be generalizable or repeatable on a corpus of similar makeup and similar or larger size. However, we can cautiously use the results as a litmus for where future research may go.

From the perspect of L1ML (the markup language), the results appear to suggest that most of the fine-grained detail encoded in various different types of errors was less useful of a metric than overall error counts in general. While this may seem to corroborate the merit of using counts and other structural features in NLI tasks, we do not think it is prudent to discount the use of annotation-based corpus linguistics just yet, particularly given the nature of our own results. We hope to use L1ML in the future on a larger corpus for a task more limited in scope to lend credence to the markup of morphosyntactic error features.

In general, the annotator experience was reported as being a difficult one. Annotators reported the difficulty of the task, the high amount of linguistic knowledge required to understand the specification, the length of the documents, and the large amount of time (about 15-20 min/document) required to annotate a specific document. In lieu of this length of time, one of our annotators reported running out of time and simply doing the documents as fast as possible, potentially leading to the lack of agreement reported for our IAA. To rectify this, we have many suggestions to propose. First, that annotators be trained during training sessions where researchers and annotators work together to annotate documents from the dev set, and then work with researchers during working sessions where researchers can be present to answer any questions or confusion. This has the potential to allow for greater IAA and standardization from the beginning of the project, given that all annotators would be present. It also allows for an easier way to push out updates and standardize quickly with these updates. Furthermore, we propose that a standardized forum to answer questions

(rather than an e-mail update) would allow questions to be answered more readily and provide a digital document in the vein of a specification for annotators to readily access. It is our hope that the implementation of these suggestions would improve IAA among our annotators.

Further research using L1ML and other corpus-based methods for NLI could explore using additional feature extraction methods and other machine learning algorithms, such as MaxEnt or SVM, to see if more accurate, state-of-the-art results could be implemented.

Acknowledgments

The authors would like to acknowledge Ariella Levine, Clay Riley, and Patricia Whitlock for their annotations of the M-TOEFL7 corpus, Keigh Rim for his work on MAE, and James Pustejovsky for introducing the authors to the world of annotation. All errors and mistakes are, of course, the responsibility of the authors.

6. Bibliographical References

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., and Chodorow, M. (2013). Toefl11: A corpus of non-native english. RR 13-14, Educational Testing Service, Princeton, New Jersey.

Dale, R., Anisimoff, I., and Narroway, G. (2012). Hoo 2012: A report on the preposition and determiner error correction shared task. In *The 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 54–62, Montréal, QC, Canada, June. Association for Computational Linguistics (ACL).

Forman, G. and Cohen, I. (2004). Learning from little: Comparison of classifiers given little training. In *8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 161–172, Pisa, Italy, September. Springer Berlin Heidelberg.

Malmasi, S. and Dras, M. (2014). Language transfer hypothesis with 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 (ACL).

Malmasi, S., Wang, S.-M. J., and Dras, M. (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 (ACL).

Tetreault, J., Blanchard, D., Cahill, A., and Chodorow, M. (2012). Native tongues, lost and found: Resources and empirical evaluations in native language identification. In *Proceedings of COLING 2012: Technical Papers*, pages 2585–2602, Mumbai, India, December. International Conference on Computational Linguistics (COLING).

Tetreault, J., Blanchard, D., and Cahill, A. (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 48–47, Atlanta, Georgia, June. Association for Computational Linguistics (ACL).

Wang, S.-M. J. and Dras, M. (2011). Exploiting parse structures for native language identi?cation. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1600–1610, Edinburgh, Scotland, UK, July. Association for Computational Linguistics (ACL).