Native Language Identification with User Generated Content

Gili Goldin
Dept. of Computer Science
University of Haifa
Israel
gili.sommer@gmail.com

Ella Rabinovich[*]
Dept. of Computer Science
University of Toronto
Canada
ellarabi@gmail.com

Shuly Wintner
Dept. of Computer Science
University of Haifa
Israel
shuly@cs.haifa.ac.il

Abstract

We address the task of native language identification in the context of social media content, where authors are highly-fluent, advanced nonnative speakers (of English). Using both linguistically-motivated features and the characteristics of the social media outlet, we obtain high accuracy on this challenging task. We provide a detailed analysis of the features that sheds light on differences between native and nonnative speakers, and among nonnative speakers with different backgrounds.

1 Introduction

The task of *native language identification* (NLI) aims at determining the native language (L1) of an author given only text in a foreign language (L2). NLI has gained much popularity recently, usually with an eye to educational applications (Tetreault et al., 2013): the errors that learners make when they write English depend on their native language (Swan and Smith, 2001), and understanding the different types of errors is a prerequisite for correcting them (Leacock et al., 2010). Consequently, tutoring applications can use NLI to offer better targeted advice to language learners.

However, the NLI task is not limited to the language of learners; it is relevant also, perhaps even more so, in the (much more challenging) context of highly-fluent, advanced nonnative speakers. While the English language dominates the internet, native English speakers are far outnumbered by speakers of English as a foreign language. Consequently, a vast amount of static and dynamic web content is continuously generated by nonnative writers. Developing methods for identifying the native language of nonnative English authors on social media outlets is therefore an important and pertinent goal.

We address the task of native language identification in the context of user generated content (UGC) in online communities. Specifically, we use a large corpus of English *Reddit* posts in which the L1 of authors had been accurately annotated (Rabinovich et al., 2018). On this dataset, we define three closely-related tasks: (i) distinguishing between native and nonnative authors; (ii) determining to which language family the native language of nonnative authors belongs; (iii) identifying the native language of nonnative authors. Importantly, we employ features that take advantage of both linguistic traits present in the texts and the characteristics of the social media outlet. We obtain excellent results: up to 92% accuracy for distinguishing between natives and nonnatives, and up to 69% for the 23-way NLI classification task.[1]

The contribution of this paper is manifold. First, this is one of the first works to address NLI with highly-advanced nonnatives; it is also among the first to address the task in the context of UGC. Furthermore, we define a plethora of features, some which have been used in earlier works but others that are novel. In particular, we define a set of features that rely on the characteristics of the social media outlet, thereby extending the task somewhat, from linguistic analysis to user profiling. Finally, we provide a detailed analysis of the results, including the specific contribution of various features and feature sets. This analysis will be instrumental for future extensions of our work.

2 Related work

The NLI task was introduced by Koppel et al. (2005), who worked on the International Corpus of Learner English (Granger, 2003), which includes texts written by students from Russia, the

[*]Work done while the second author was at the University of Haifa.

[1] It is important to note that some of our features are specific to the Reddit corpus and will not easily generalize to other datasets.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3591–3601
Brussels, Belgium, October 31 - November 4, 2018.

Czech Republic, Bulgaria, France, and Spain. The same experimental setup was adopted by several other authors (Tsur and Rappoport, 2007; Wong and Dras, 2009, 2011). The task gained popularity with the release of nonnative *TOEFL* essays by the Educational Testing Service (Blanchard et al., 2013); this dataset has been used for the first NLI Shared Task (Tetreault et al., 2013) and also for the 2017 NLI Shared Task (Malmasi et al., 2017).

Our task is closely related to the task of *dialect identification*, in which the goal is to discriminate among similar languages, language varieties and dialects. Classic machine learning classification methods are usually applied for this task, often with SVM models. The best reported features include word and character n-grams, part of speech n-grams and function words (Zampieri et al., 2017; Malmasi and Zampieri, 2017).

The current state of the art in NLI, according to Malmasi and Dras (2017), utilizes some variant of contemporary machine learning classifier with the following types of features: (i) word, lemma and character n-grams, (ii) function words (FW), (iii) part-of-speech (POS) n-grams, (iv) adaptor grammar collocations, (v) Stanford dependencies, (vi) CFG rules, and (vii) Tree Substitution Grammar fragments. The best result under cross-validation on the TOEFL dataset, which includes 11 native languages (with a rather diverse distribution of language families), was 85.2% accuracy. Applying these methods to different datasets (the ASK corpus of learners of Norwegian (Tenfjord et al., 2006) and the Jinan Chinese Learner Corpus (Wang et al., 2015), 10-11 native languages in each) resulted in 76.5% accuracy for the Chinese data and 81.8% for the Norwegian data, with LDA-based classification yielding top results.

Notably, all these works identify the native language of *learners*. Identifying the native language of advanced, fluent speakers is a much harder task. Furthermore, our dataset includes texts by native speakers of 23 languages, more than double the number of languages used in previous works; and our L1s are all European, and often typologically close, which makes the task much harder.

Two recent works address the task of NLI on UGC in social media. Anand et al. (2017) summarized the shared task on Indian NLI: given a corpus of Facebook English comments, the task was to identify which of six Indian languages is the L1 of the author. The best reported result was 48.8%, obtained by an SVM with character and word n-grams as features. These are content based features that are highly domain-dependent and are not likely to generalize across domains. Volkova et al. (2018) explored the contribution of various (lexical, syntactic, and stylistic) signals for predicting the foreign language of non-English speakers based on their English posts on Twitter. This effectively results in a 12-way classification task, with 12 different L1s (data sizes are distributed very unevenly), and the best results are unsurprisingly obtained with word unigrams and bigrams.

In contrast to these two studies, we work with many more L1s (23); we explore various types of features, including features based on social network structures and content-independent features; and we evaluate our classifiers both in and outside of the domain of training.

Several works address social aspects of social networks, and in particular identify “influential” users (Ghosh and Lerman, 2010; Trusov et al., 2010; Afrasiabi Rad and Benyoucef, 2011). Network structure has been shown to be useful in other tasks of user profiling, such as geolocation (Jurgens et al., 2015). Our design of the social network features (Section 3.5.4) are motivated by these works.

Works that aim to distinguish between native and nonnative authors (Bergsma et al., 2012; Rabinovich et al., 2016; Tomokiyo and Jones, 2001) typically rely on lexical and grammatical characteristics that reflect influences of L1 on L2. We used such features, but augmented them by features that can be induced from the network structure of social media outlets (Jurgens et al., 2015). To the best of our knowledge, ours is the first work that extensively exploits social network properties for the task of NLI. Our work is also inspired by research on the (related but different) task of identifying translations (Baroni and Bernardini, 2006; Rabinovich and Wintner, 2015; Volansky et al., 2015) and their source language (Koppel and Ordan, 2011; Rabinovich et al., 2017).

3 Experimental setup

3.1 Dataset

Reddit is an online community consisting of thousands of forums for news aggregation, content rating, and discussions. Content entries are organized by areas of interest called *subreddits*, ranging from main forums that receive much attention to smaller ones that foster discussion on niche areas. Subreddit topics include news, science, arts, and many others. An increasing body of work has used Reddit data for social media analysis (Jurgens et al., 2015; Newell et al., 2016, and many more).

We used the Reddit dataset released by Rabinovich et al. (2018). It includes Reddit posts (both initial submissions and subsequent comments), focusing on subreddits (Europe, AskEurope, EuropeanCulture) whose content is generated by users specifying their country as a *flair* (metadata attribute). We refer to these subreddits as *European*. Following Rabinovich et al. (2018), we view the country information as an accurate, albeit not perfect, proxy for the native language of the author.

Rabinovich et al. (2018) justified their trust in the accuracy of the L1 annotation; we conducted an additional validation of the data.[2] We used a specific Reddit thread in which users were asked to comment in their native language. We collected the comments in this thread of all the users in our dataset. Then, we used the Polyglot language identification tool to determine the language of the comments. We filtered out short comments, comments for which the tool's confidence was low, and comments in English of users from non-English speaking countries. Of the remaining 572 users, 479 (84%) contributed comments in the language that we considered their native. We inspected the remaining users, and for many (albeit not all) we attribute the mismatch to errors in the tool (i.e., comments in Serbian written in the Latin alphabet are wrongly predicted to be in closely-related Slavic languages). We conclude that the accuracy of the L1 annotation is high; finally, we note additionally that any noise in this labeling can only work against us in this work.

We filtered out data from multilingual countries (Belgium, Canada, and Switzerland). Rabinovich et al. (2018) showed that the English of reddit non-native authors is highly advanced, almost at the level of native speakers, making the NLI task particularly demanding.

All the posts in the dataset are associated with a unique user ID. The dataset also contains submissions and comments in other subreddits that were written by the same authors, based on their user ID. This provided us with an out-of-domain test set for evaluating the robustness of our methods. The collected data reflect about 50 (mostly European) countries, and consist of over 230M sentences, or 3.5B tokens, annotated with authors' L1.

3.2 Preprocessing

Each sentence in the dataset is tagged with the author's user ID, the subreddit it appeared in and the author's country. We segmented the dataset into *chunks* of 100 sentences, each chunk containing sentences authored by the same user.[3] The sentences were kept in their original order in the posts; users with fewer than 100 sentences were filtered out. We also left out native languages with fewer than 100 users (after the initial filtering). The resulting dataset includes 23 native languages spanning 29 countries, and consists of 34,511 unique users, almost 200M sentences and over 3B tokens. The countries and languages reflected in the dataset are listed in Table 10 (see Supplementary Materials).

All chunks were annotated for part-of-speech using Spacy. We used Aspell to spell-check the texts; every misspelled word in the original chunk was annotated with the first correction suggested by the spell checker. We also extracted from reddit additional social network properties such as users' *karma* scores, number of comments and submissions, number of comments per submission, as well as the number of months each user was active on Reddit and all the subreddits that each user in our dataset posted in (see Section 3.5.4). The processed dataset will be made publicly available.

3.3 Task

We define three classification tasks: **binary classification,** distinguishing between native and non-native authors; **language family classification** determining the language family (Germanic, BaltoSlavic, Romance, or native English) of the user; and **language identification** whose goal is to identify the native language of the user.

Different countries which have the same official language (e.g., Germany and Austria) were tagged with the same language label. For example, USA, UK, Ireland, New Zealand and Australia were all tagged with the label 'English' for the NLI task.

We then randomly downsampled the data to ensure that each class had the same number of users.

[2] We thank an anonymous reviewer for suggesting this.

[3] Similar classification tasks, e.g., the TOEFL task (Tetreault et al., 2013; Malmasi et al., 2017), used single essays as the unit for classification. Tasks aiming to identify translation and its source language typically use chunks of 2000 tokens (Volansky et al., 2015). We plan to experiment also with smaller chunks.

See the precise details in Section A.1 of the Supplementary Materials.

3.4 Methodology

We cast NLI as a supervised classification task and used *logistic regression* (as implemented in Scikit-learn) as a classification model. We defined several features that had been proven useful for similar tasks; some of them are general stylistic features that are presumably content-independent: these include function words, POS n-grams, simplification measures such as sentence length, etc. (Rabinovich and Wintner, 2015; Volansky et al., 2015). Other features are content based; most obviously, token n-grams, but also character n-grams (Avner et al., 2016). We expect content-based features to be highly accurate but also highly domain-dependent, and in the case of our dataset, topic-dependent. Content-independent features are expected to be weaker yet more robust.

In addition, we used features that reflect spelling and grammar errors. We assume that native and nonnative speakers make different kinds of errors in English, and that the errors of nonnatives may reveal traces of their L1 (Kochmar, 2011; Berzak et al., 2015).

Aiming to enhance the quality of classification we exploited properties that can be induced from conversational networks. We hypothesize that native speakers of the same language tend to interact more with each other (than with speakers of other languages). We hypothesize further that native speakers post more than nonnatives, and hence we defined user *centrality* measures that reflect that. We also hypothesize that native speakers' posts tend to be more spontaneous, coherent and clear, thereby drawing more attention. To reflect that, we counted the number of comments, upvotes and down-votes that were submitted to each post. While these and similar properties have been studied in the domain of social networking, to the best of our knowledge this is the first attempt to use an extensive set of features inferred from social networks for the NLI task.

3.5 Features

We designed several features to be used in all three tasks. In this section we describe these features.

3.5.1 Content features

Authors are more likely to write about topics that are related to their country and their culture, hence features that reflect content may help distinguish among authors from different countries (and, therefore, languages). For example, the word *'Paris'* is more likely to occur in texts written by French authors, while the word *'canal'* is more likely to occur in texts of Dutch authors. We defined features that take text content into account. We expect these features to yield high accuracy when testing on the training domain, but much lower accuracy when testing on different domains.

Character tri-grams The top 1000 most frequent character 3-grams in the dataset were used as features. For each chunk the value of a certain character 3-gram feature was the number of its occurrences in the chunk normalized by the total number of character 3-grams in the chunk.

Token uni-grams The top 1000 most frequent tokens in the dataset were used as features. For each chunk the value of a certain token feature was the number of its occurrences in the chunk normalized by the total number of tokens in the chunk.

3.5.2 Spelling and grammar

We used a spell checker (Section 3.1) to discover the (first) closest correction for each word marked as incorrect. Based on this correction, we defined several edit-distance-based features using Python's Python-Levenshtein extension.

Edit distance Assuming that nonnative speakers will make more spelling errors than natives, we used the average Levenshtein distance between the original word and the correction offered by the spell checker, for all words in a chunk, as a feature.

Spelling errors Again, we assume that the spelling errors that nonnatives make may reflect properties of their L1; this has already been shown for learners (Tsvetkov et al., 2013). Using the edit distance between a mis-spelled word w in a text chunk, marked by the spell checker, and its suggested correction c, we extract insertions, deletions and substitutions that yield c from w and use them as features. For each chunk, the value of this feature is the number of occurrences of each substitution (a character pair), insertions, and deletions in the chunk. We only used the top-400 most frequent substitutions.

We initially classified spelling errors as content-independent features, assuming that they would reflect transfer of linguistic phenomena from L1. However, having analyzed this feature type, we observed that many of the mis-spelled words turned out to be non-English words, which apparently are abundant in our dataset even after removing non-English sentences. We therefore view this feature as content dependent.

Grammar errors We hypothesize that grammar errors made by nonnatives may reflect grammatical structures revealing their L1. We therefore used LanguageTool, a rule-based grammar checker, to identify grammatical errors in the text.[4] We defined an indicator binary feature for each of the (over 2000) grammar rules detected by the grammar checker.[5]

3.5.3 Content-independent features

Content-based features may overly depend on the domain of the training data, and consequently be less effective when testing on different domains. Content-independent features are expected to be more robust when they are used out-of-domain.

Function words Function words are highly frequent and as such they are assumed to be selected unconsciously; they are therefore considered to reflect style, rather than content. Function words have been used successfully in a variety of style-based classification tasks (Mosteller and Wallace, 1963; Koppel and Ordan, 2011; Volansky et al., 2015; Rabinovich et al., 2016). We used as features (the frequencies of) about 400 function words, taken from Volansky et al. (2015).

POS tri-grams POS n-grams are assumed to reflect (shallow) grammar. The native language of the author is likely to influence the structure of his or her productions in English, and we assume that this will be reflected in this feature set. We used as features the normalized frequency of the top 300 most frequent POS tri-grams in the data set.[6]

Sentence length Texts of nonnative speakers are assumed to be simpler than those of natives; in particular, we expect them to have shorter sentences. The value of this feature is the average length of the sentences in the chunk.

3.5.4 Social network features

We defined several features that are extracted from the social network data, particularly its structure. First, we defined feature sets that express the *centrality* of users, under the assumption that native speakers would be more central on social networks. Consequently, this set of features is expected to be beneficial mainly for the binary native/nonnative classification.

User centrality in the social network of Reddit can be reflected in various ways:

Karma Reddit assigns a *karma* score to each user. This score “reflects how much good the user has done for the reddit community. The best way to gain karma is to submit links that other people like and vote for”.[7] The Karma score is an undisclosed function of two separate scores: *link karma*, which is calculated from the user's posts that contain links, and *comment karma*, which is computed from the user's comments. We extracted both types of karma scores for all users in the dataset and used each of them (specifically, the user's monthly average scores) as a feature.

Average score Reddit calculates a *score* for each submission as the number of up-votes minus the number of down-votes the submission received. We used the user's average score per month as a feature.

Average number of submissions We counted for each user the total number of submissions he or she authored. For each chunk the value of this feature is the user's average number of submissions per month.

Average number of comments Same as the above, but counting user's comments (responses to submissions) instead of submissions.

Most popular subreddits Finally, we assume that native speakers of the same language tend to interact more with each other than with others, and we also assume that they are more likely to be interested in similar topics, influenced by their country and culture; specifically, we hypothesize that the forums in which users post most will be common for users from the same country. Therefore, we extracted for each country in the dataset the most popular subreddits among users from this country. For each country, we sorted subreddits according to the number of users from this country who posted at least once in this subreddit. The 30 most popular subreddits of each country

[4] We used the Python wrapper for LanguageTool.

[5] The list of English grammar rules is available online.

[6] We also experimented with POS 5-grams but they did not yield better results.

[7] The Reddit Wiki.

were taken as features. The unique list of popular subreddits contains 141 subreddits. For each chunk the value of a certain subreddit feature was a binary value indicating whether or not the author of this chunk has posted in this subreddit.

3.6 Evaluation

It is well known that similar classification tasks are highly domain-dependent; simply put, the properties of the domain overshadow the much more subtle signal of the author's L1. To test the robustness of various feature sets in the face of domain noise, we defined two evaluation scenarios: *in-domain*, where training and testing is done only on chunks from the European subreddits; and *out-of-domain*, where we train on chunks from the European subreddits and test on chunks from other subreddits, making sure they were authored by different users. Note that the out-of-domain corpus spans tens of thousands of subreddits with a huge number of topics. The precise evaluation scenario is somewhat involved and is detailed in Section A.2 of the Supplementary Materials. We report *accuracy*, defined as the percentage of chunks that were classified correctly out of the total number of chunks.

4 Results

We implemented the features discussed in Section 3.5 and evaluated the accuracy of the three classification tasks mentioned in Section 3.4 under the configurations described in Section 3.6. The trivial baseline for the binary classification task is 50%, for language family classification 25%, and for the language identification task 4.35%.

4.1 Individual feature sets

The accuracy results for each feature set described in Section 3.5 for the in-domain evaluation scenario are presented in Table 1.

Feature Set	Binary	Families	NLI
Char. 3-grams	85.58	78.20	62.06
Token unigrams	86.26	69.36	31.26
Spelling	71.04	52.18	27.74
Grammar errors	66.79	37.96	8.36
FW	80.34	57.80	20.15
POS 3-grams	69.14	50.29	13.30
Sentence length	50.37	26.14	4.79
Social network	57.92	32.39	5.75
Subreddits	87.08	82.56	74.46

Table 1: In-domain accuracy, individual feature sets

Evidently, all feature sets outperform the baseline, although some are far better than others. The feature that yields the best accuracy is *Subreddits*, with 87% accuracy on the binary task, 82% on the language family task and 74% on the NLI task. We elaborate on this feature in Section 4.2 below. As expected, the content based features yield relatively high results when the evaluation is in-domain. POS 3-grams and function words yield reasonable results, but not as good as in other classification setups (e.g., Rabinovich et al. (2016)), where the evaluation was done by shuffling texts of various users. As we evaluate on chunks of single users, the personal style of the user may dominate the subtler signal of his or her native language. Sentence length performs poorly, even on the binary task. Our assumption was that the *social network* feature set will work well only for the binary classification; this seems to be borne out by the results.

4.2 Feature combination

We now set out to investigate different feature combinations in both evaluation scenarios, aiming to define feature types that yield the best in-domain accuracy, as well as those that are most robust and generalize well out-of-domain.

Table 2 depicts the results obtained by combining character trigrams, tokens, and spelling features (Sections 3.5.1, 3.5.2). As expected, these content features yield excellent results in-domain, but the accuracy deteriorates out-of-domain, especially in the most challenging task of NLI.

	Binary	Families	NLI
In-domain	91.07	83.51	70.26
Out-of-domain	81.49	65.37	35.99

Table 2: Results: content features

The content-independent features (Section 3.5.3), whose contribution is depicted in Table 3, indeed fare worse, but are seemingly more robust outside the domain of training.

	Binary	Families	NLI
In-domain	81.89	62.40	22.38
Out-of-domain	74.56	52.35	14.86

Table 3: Results: content-independent features

Table 4 shows the results obtained by combining the spelling features with the grammar features

(Section 3.5.2). Clearly, these two feature types reflect somewhat different phenomena, as the results are better than using any of the two alone.

	Binary	Families	NLI
In-domain	72.93	55.59	26.74
Out-of-domain	70.24	47.23	14.15

Table 4: Results: grammar and spelling features

Table 5 shows the accuracy obtained by all the centrality features (Section 3.5.4), excluding the most popular subreddits. As expected, the contribution of these features is small, and is most evident on the binary task. The signal of the native language reflected by these features is very subtle, but is nonetheless present, as the results are consistently higher than the baseline.

	Binary	Families	NLI
In-domain	57.92	32.39	5.75
Out-of-domain	56.29	30.70	5.60

Table 5: Results: centrality features

Finally, the contribution of the most popular subreddits feature is shown in Table 6. The results for this single feature type are superb, both in- and out-of-domain. However, as this feature is unique to the dataset used for the present work, it is hard to see it generalized to similar tasks that use other datasets, even in the context of UGC.

	Binary	Families	NLI
In-domain	87.08	82.56	74.46
Out-of-domain	85.49	82.17	73.63

Table 6: Results: most popular subreddits

Therefore, we report the results obtained with *all* features, with (Table 7) and without (Table 8) the reddit-specific most popular subreddit feature.

	Binary	Families	NLI
In-domain	93.40	90.41	86.05
Out-of-domain	87.19	83.43	78.99

Table 7: Results: all features

	Binary	Families	NLI
In-domain	92.21	82.51	68.97
Out-of-domain	79.34	66.21	36.16

Table 8: Results: all features except subreddits

Summing up, we have shown that the challenging task of native language identification in the context of user generated content, where English texts are authored by highly competent nonnative speakers with as many as 23 native languages, can be accomplished with very high accuracy, as high as 86% when evaluated in-domain, and almost 79% out-of-domain (Table 7). While these results deteriorate when the specific characteristics of our dataset are not taken advantage of, we still obtain very high accuracy on the binary task of distinguishing native from nonnative speakers, and on the four-way task of identifying the language family of the authors' L1 (Table 8).

4.3 Dialect robustness

To assess the robustness of our results, especially in the context of dialect identification, we repeated the experiments in a special scenario: we trained classifiers on all the data, but removed from the English training set users from Ireland. Then, we tested the classifiers only on users from Ireland. We used all the features listed above, except the subreddit feature.

The results are 59.09% accuracy in-domain, compared with 69.21% in the standard scenario, where users from Ireland are also used for training; and 37.51% out-of-domain, compared with 47.85% in the standard scenario. In both cases, accuracy drops by 10 percent points. We conclude that our method is reasonably robust to dialectal variation, at least in the case of English varieties.

5 Analysis

We now set out to analyze some of the more interesting features, both in terms of their contribution to the accuracy of the classification and in terms of what they reveal about the English of advanced nonnative speakers.

5.1 Social network features

Subreddits This feature set works so well because many of the most popular subreddits in which users post are culturally revealing. Specifically, there is a significant presence in this list to (subreddits focusing on) specific countries. Very likely, most of the active users in those subreddits reside in these countries, thereby revealing their native language. This corroborates our hypothesis that native users of the same language tend to be active in mutual subreddits.

Network structure Table 9 lists the average values of the centrality features, comparing native vs. nonnative authors. The average values are higher for native users than for the nonnative ones in all of the centrality features, as we hypothesized. Evidently, native speakers are more central in social networks than nonnative ones.

	Native		nonnative	
	Avg	std	Avg	std
Score	1349	2383	906	1886
# comments	147	173	92	112
# submissions	5	21	4	13
Comment karma	787	1260	529	837
Link karma	202	1012	141	580

Table 9: Centrality features: average values and standard deviation

5.2 Spelling

Edit Distance As expected, the average word edit distance of native users (0.048) was significantly lower compared to nonnative ones (0.071).

Substitutions Most revealing was the analysis of substitutions suggested by the spell checker, as they shed light on phonetic and orthographic influences of the authors' L1 on their English. We list below some of the most common spelling errors.

Vowels Replacing *'e'* with *'a'* was twice as common among nonnative users than native ones. Examples include *'existance'*, *'independance'*, *'privillages'*, and *'apparantly'*. Similarly, replacing *'y'* with *'i'* was three times more common for nonnatives: *'synonims'*, *'analized'*, etc. Replacing *'o'* with *'a'* was common among nonnatives, especially in the context of diphthongs: *'enaugh'* instead of *'enough'*, or *'cauntry'* for *'country'*.

Voicing Replacing *'f'* with *'v'* was common mostly among German speakers: *'devense'*, *'bevore'*, *'sacrivice'*, etc. Another error that was relatively common in texts written by German speakers is the replacement of *'d'* with *'t'*: *'unterstand'*, *'canditate'*, *'upgradet'*, *'hundret'*, etc. Confusing *'z'* with *'s'* was very common across all L1s, even for natives. Among native users this reflects spelling variations between US and UK English. Thus, the spell-checker marks the following forms, i.a., in New Zealand English: *'Organisation'*, *'Recognise'*, *'Realise'*, *'Criticise'*, etc. Replacing *'s'* with *'z'* was not as common in the dataset, and was present mostly in texts of French users: *'advertize'*, *'tablez'*, and, most frequently, *'surprize'*.

Other substitutions Replacing *'c'* with *'k'* was almost four times more common with nonnatives; it was significantly more common among Germanic and Balto-Slavic speakers, and much less common among Romance speakers. Examples include *'inspektor'*, *'klassik'*, etc. Replacing *'t'* with *'c'* was common in words in which the *'t'* is pronounced [ʃ]: *'negociate'*, *'nacional'*. This error was prevalent in texts of Spanish authors.

Insertions and deletions Insertion of *'o'* was common for all nonnative speakers, often when the word contains one *'o'* but the pronunciation is [u], e.g., *'proove'* instead of 'prove'. Spurious occurrences of *'e'* were also very common among all nonnative users, especially authors whose L1 was French: *'gover**n**ement'*, *'unemploy**e**ment'*, *'explicit**e**ly'*. Deletions of *'e'* were also very common, especially in the context of words that end with *'ely'*: *'definitly'*, *'completly'*, *'extremly'*, *'absolutly'*, etc. Spurious instances of *'u'* were mostly present in texts of authors with Germanic and Romance L1s, e.g.: *'languague'*, *'percentuage'*.

Wrong insertions of *'l'* were very common, especially at the end of words that end with *'l'*: *'untill'*, *'controll'*, *'usefull'*. Deletion of *'l'* was common for all nonnative users, especially with Balto-Slavic L1s. The most common context for this error is words ending with *'ally'*: *'literaly'*, *'actualy'*, *'basicaly'*, *'illegaly'*, *'totaly'*, *'personaly'*, etc.

The most common deletion among nonnatives was omission of the first *'r'* in *'surprise'*, followed by omitting the first *'n'* in *'government'*.

5.3 Grammar

We list below some of the grammar rules whose violations distinguish well between native and nonnative speakers, using the original grammar checker rule names. Unsurprisingly, several grammar rules were violated much more (twice as frequently) by nonnative users:

adverb_word_order wrong position of adverb, e.g., *'people sometimes will respond'* instead of *'people will sometimes respond'*.

cd_nn agreement error of a numeral followed by a singular count noun, e.g., *'I have 5 book'*.

this_nns using *'this'* instead of *'these'* or vice versa, e.g., *'you don't know what these symbol* *represent'*.

did_baseform using a tensed verb after *'did'* or *'didn't'*: *'the court didn't gave him a fair trial'*.

a_uncountable an indefinite article before non-count nouns: *'smaller places have an access to...'*.

fewer_less confusing *'fewer'* with *'less'*: *'with less possibilities'*.

much_countable using *'much'* instead of *'many'*: *'no matter how much people'*. This error was much more common among nonnative users, although, among native speakers, it was significantly more common in texts written by users from New-Zealand and Ireland than in texts of other English speaking users.

en_a_vs_an confusing *'a'* with *'an'*: *'it provides a organized way to discuss'*. This error was very common among speakers of Germanic and Romance languages, but less common among speakers of Balto-slavic languages (presumably due to the lack of articles in their L1s).

In contrast, some grammar rules were violated more by native speakers:

possessive_apostrophe omitting the apostrophe in possessive *''s'*: *'they had 20% of the worlds remittance'*. This error was more than twice as common in texts of natives.

try_and the verb *'try'* followed by *'and'*; this is common in colloquial speech, but is prescriptively wrong: *'a candidate should try and represent'*. This rule was violated over three times more frequently by native speakers (but rarely in texts of New-Zealand users).

their_is *'there'* and *'their'* are commonly confused; this rule spots such cases by the presence of *'be'*: *'their are a lot of'*.

about_its_nn confusing *'its'* and *'it's'* is common; this rule identifies wrong usage after a preposition: *'lash out regularly towards it's neighbors'*. This error was most common in texts of English speakers from Australia, Ireland and the UK, but not the US.

Summing up, it seems that nonnative speakers make more grammatical errors, while the mistakes of native speakers either stem from sloppy writing style and lack of attention, or reflect style variations and casual style rather than actual errors.

6 Conclusion

We described a system that can accurately identify the native language of highly-advanced, fluent nonnative authors as reflected in the social media Reddit corpus. This is among the first studies to perform NLI in the highly challenging scenario of user generated content, particularly at such a large scale. We showed that while content-dependent features yield more accurate results, features that abstract away from content tend to be more robust when tested out of the domain of training. The in-depth analysis of spelling and grammar errors demonstrates that mistakes made by nonnative speakers reflect traces of their native language. We also illuminated some of the social characteristics of native and nonnative authors on social media outlets.

Our future plans include adaptation of the trained models to additional corpora, e.g., user generated content collected from *Facebook* and *Twitter*. Furthermore, we plan to devise *unsupervised* approaches to the identification of native language with the same dataset. We would also like to test the classifiers defined here in the more challenging scenario of smaller text chunks (e.g., 10–20 sentences rather than the 100-sentence text chunks we used here). Finally, we are currently experimenting with adversarial learning models for this task.

Acknowledgements

This research was supported by Grant No. 2017699 from the United States-Israel Binational Science Foundation (BSF) and by Grant No. 1813153 from the United States National Science Foundation (NSF).

References

Amir Afrasiabi Rad and Morad Benyoucef. 2011. Towards detecting influential users in social networks. In *E-Technologies: Transformation in a Connected World*, pages 227–240, Berlin, Heidelberg. Springer.

Kumar M Anand, Ganesh HB Barathi, Shivkaran Singh, KP Soman, and Paolo Rosso. 2017. Overview of the INLI PAN at FIRE-2017 track on Indian native language identification. Unpublished manuscript.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, 31(1):30–54.

Marco Baroni and Silvia Bernardini. 2006. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Shane Bergsma, Matt Post, and David Yarowsky. 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.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2015. Contrastive analysis with predictive power: Typology driven estimation of grammatical error distributions in ESL. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 94–102.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Rumi Ghosh and Kristina Lerman. 2010. Predicting influential users in online social networks. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*.

Sylviane Granger. 2003. The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, pages 538–546.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, pages 188–197.

Ekaterina Kochmar. 2011. Identification of a writer's native language by error analysis. Master's thesis, University of Cambridge.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Automatically determining an anonymous author's native language. *Intelligence and Security Informatics*, pages 41–76.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan and Claypool.

Shervin Malmasi and Mark Dras. 2017. Native language identification using stacked generalization. ArXiv:1703.06541 [cs.CL].

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2017. Arabic dialect identification using iVectors and ASR transcripts. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 178–183. Association for Computational Linguistics.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309.

Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on reddit during a period of community unrest. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media*, pages 279–288.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1870–1881.

Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540. Association for Computational Linguistics.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Translactions of the Association for Computational Linguistics*, 6.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised identification of translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Michael Swan and Bernard Smith. 2001. *Learner English*, second edition. Cambridge University Press, Cambridge.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

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 Building Educational Applications Using NLP. Association for Computational Linguistics.

Laura Mayfield Tomokiyo and Rosie Jones. 2001. You're not from 'round here, are you?: naive Bayes detection of non-native utterance text. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics.

Michael Trusov, Anand V. Bodapati, and Randolph E. Bucklin. 2010. Determining influential users in internet social networks. *Journal of Marketing Research*, 47(4):643–658.

Oren Tsur and Ari Rappoport. 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.

Yulia Tsvetkov, Naama Twitto, Nathan Schneider, Noam Ordan, Manaal Faruqui, Victor Chahuneau, Shuly Wintner, and Chris Dyer. 2013. Identifying the L1 of non-native writers: the CMU-Haifa system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 279–287. Association for Computational Linguistics.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Svitlana Volkova, Stephen Ranshous, and Lawrence Phillips. 2018. Predicting foreign language usage from English-only social media posts. In *Proceedings of NAACL-2018*.

Maolin Wang, Shervin Malmasi, and Mingxuan Huang. 2015. The Jinan Chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 118–123.

Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop 2009*, pages 53–61, Sydney, Australia.

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. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15. Association for Computational Linguistics.