Native Language Identification with User Generated Content

Gili Goldin

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa
Faculty of Social Sciences
Department of Computer Science

August 2019

Native Language Identification with User Generated Content

By: Gili Goldin

Supervised by: Prof. Shuly Wintner

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE MASTER'S DEGREE

University of Haifa
Faculty of Social Science
Department of Computer Science

August 2019

Approved by: _____ Date: _____
(Supervisor)

Approved by: _____ Date: _____
(Chairperson of Master's studies Committee)

Acknowledgments

I would first like to greatly thank my thesis advisor Prof. Shuly Wintner for his valuable supervision and encouragement. Shuly has exposed me to the fascinating world of natural language processing and to the beauty of linguistics. Thanks to Shuly I was able to broaden my knowledge and experience with many challenges throughout my research. Shuly constantly motivated me to strive for more. His guidance and directions were always very helpful while still allowing this paper to be my own work.

I wish to express a special gratitude to Dr. Ella Rabinovich for her insightful advices and great ideas. Ella was always very supportive and helpful. It was a great pleasure working with her.

I would also like to thank Hila Rozenberg for her kindest help and patience in answering any administrative question and for guiding me through the university's bureaucracy.

I am grateful to my family for all their support and their faith in me. A special thanks to my parents who have been a great source of inspiration for me and raised me to be inquisitive, ambitious and to believe in my abilities.

Last, and most important, my deepest gratitude to my dear husband Dima for his love, patience, encouragement and support. I couldn't have done it without you.

Contents

Native Language Identification with User Generated Content

Gili Goldin

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. Lastly, we use neural networks for this task in order to boost the accuracy and achieve state of the art results.

List of Tables

List of Figures

Chapter 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 80.7% for the 30-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. Additionaly, 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. Finaly, we use neural networks and explore different word embeddings in order to improve the results and make the work comparable to other recent works.

[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.

Chapter 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 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 (Malmasi and Zampieri, 2017; Zampieri et al., 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 30 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) did not address NLI directly, but 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 (30); 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 (Afrasiabi Rad and Benyoucef, 2011; Ghosh and Lerman, 2010; Trusov et al., 2010). 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 3.4.3) 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 (Avner et al., 2016; Baroni and Bernardini, 2006; Rabinovich and Wintner, 2015; Volansky et al., 2015) and their source language (Koppel and Ordan, 2011; Rabinovich et al., 2017).

Chapter 3

Experimental setup

We define three classification tasks:

binary classification, distinguishing between native and nonnative authors;

language family classification determining the language family (Germanic, BaltoSlavic, Romance, native English, or Other) of the user; and

language identification whose goal is to identify the native language of the user.

3.1 Dataset

Reddit is an online community consisting of thousands of forums for news aggregation, content rating, and discussions. As of July 2019, Reddit is ranked the fifth most visited website in the United States, and13th in the world, with over 330 million users. 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). Following Rabinovich et al. (2018), we view the country information as an accurate, albeit not perfect, proxy for the native language of the author. We refer to these

subreddits, and the texts extracted from them, as *European*. All the posts in the dataset are associated with a unique user ID; using the European dataset as a seed, we extracted all the submissions and comments of the users included in it from *all* other subreddits in reddit. We refer to these other subrreddits, and the texts included in them, as *non-European*. In this work, we use the European dataset as in-domain data, whereas the non-European one, although it is authored by the same Reddit users, is considered out-of-domain, and is used for evaluating the robustness of our methods The collected data reflect about 50 (mostly European) countries,[1] and consist of over 230M sentences, or 3.5B tokens, annotated with authors' L1.

Rabinovich et al. (2018) justified their trust in the accuracy of the L1 annotation; we conducted an additional validation of the data. 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.

Rabinovich et al. (2018) showed that the English of reddit nonnative authors is highly advanced, almost at the level of native speakers, making the NLI task particularly demanding.

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. 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. The countries and languages reflected in the dataset are listed in Table 3.1.

[1] We filtered out data from multilingual countries (Belgium, Canada, and Switzerland).

Country	Language	Country	Language
Albania	Albanian	Iceland	Icelandic
Austria	German	Italy	Italian
Germany	German	Latvia	Latvian
Australia	English	Lithuania	Lithuanian
Ireland	English	Netherlands	Dutch
New Zealand	English	Norway	Norwegian
United Kingdom	English	Poland	Polish
United States	English	Portugal	Portuguese
Bosnia	Bosnian	Romania	Romanian
Bulgaria	Bulgarian	Russia	Russian
Croatia	Croatian	Serbia	Serbian
Czech Republic	Czech	Slovakia	Slovak
Denmark	Danish	Slovenia	Slovenian
Estonia	Estonian	Spain	Spanish
Finland	Finish	Mexico	Spanish
France	French	Sweden	Swedish
Greece	Greek	Turkey	Turkish
Hungary	Hungarian	Ukraine	Ukrainian

Table 3.1: Countries and Languages in Dataset

We segmented the dataset into *chunks* of 100 sentences, each chunk containing sentences authored by the same user. 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 50 users (after the initial filtering). The resulting dataset includes 30 native languages spanning 36 countries, and consists of 39,544 unique users, almost 225M sentences (8,291,600 in-domain and 215,865,300 out-of-domain) and over 3B tokens. We then randomly downsampled the data to ensure that each class had the same number of users. To do so, we calculated the number of users tagged with each label in the data (the label can be a language, a family, or native/nonnative, depending on the task). We then randomly selected the minimum number of users with each label. Note that the number of chunks per label is still not equal because each user may have a different number of chunks; in order to cancel the bias caused by users that are over-represented in the texts of their country (i.e., users authoring a significant portion of their country's sentences), we used at most the median number of randomly selected chunks for each user. For the in-domain chunks the median is 3, for the out-of-domain ones it is 17. The median was calculated separately over all the in-domain chunks and over all the out-of domain chunks. For the out-of-domain test set we used only the out-of-domain chunks of 10% of the

users, guaranteeing that these users are disjoint of the ones in the in-domain training set; see also Section 3.5. After downsampling, we were left with 1770 unique users, 589,500 sentences (353,600 in-domain and 235,900 out-of-domain), and about 9M tokens.

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, including the users' *karma* scores, number of comments and submissions, number of comments per submission, the number of months each user was active on Reddit, and all the subreddits that each user in our dataset posted in (see Section 3.4.3). The processed dataset is publicly available.

3.3 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 (Berzak et al., 2015; Kochmar, 2011).

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, up-votes 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.

3.4 Features

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

3.4.1 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.

Spelling

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.

3.4.2 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 (Koppel and Ordan, 2011; Mosteller and Wallace, 1963; Rabinovich et al., 2016; Volansky et al., 2015). 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.[2]

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.

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.[3] We defined an indicator binary feature for each of the (over 2000) grammar rules detected by the grammar checker.[4]

3.4.3 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:

[2] We also experimented with POS 5-grams but they did not yield better results.
[3] We used the Python wrapper for LanguageTool.
[4] The list of English grammar rules is available online.

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".[5] 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 were taken as features. The unique list of popular subreddits contains 141 subreddits.

[5] The Reddit Wiki.

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.

3.5 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. In both cases, we made sure the chunks in the train set were authored by different users than those in the test set. 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 below. We report *accuracy*, defined as the percentage of chunks that were classified correctly out of the total number of chunks.

In both evaluation scenarios, a *fold* is defined over *users*, rather than text chunks. Consider first the in-domain scenario. We only consider chunks in the European subreddits, of which there are 82,916 (after downsampling). The number of users in this dataset is 39,544, but to avoid bias, we only select the minimum number of users for each label; for the NLI task, this number is 59, so we are left with $59 \times 30 = 1770$ users and 3,536 chunks. We now run 10-fold cross-validation evaluation on the set of (chunks authored by) these users, where in each fold we train on 90% of the users and test on the remaining 10%. We use the same evaluation strategy for the two other tasks.

In the out-of-domain scenario we use the much larger non-European corpus for the test set. We begin with over 2M text chunks authored by almost 40K users, but downsampling reduces this number to about 30k chunks. As above, we are left with 1700 users. We randomly select 10% of these users in a stratified way (uniformly across L1s), and use their non-European chunks for testing. For training, we use the European chunks authored by the remaining 90% users. We repeat this process ten times and report the average of the ten runs. Again, we use the same evaluation strategy for the two other tasks.

Chapter 4

Results

We implemented the features discussed in Section 3.4 and evaluated the accuracy of the three classification tasks mentioned in Section 3.3 under the configurations described in Section 3.5. The trivial baseline for the binary classification task is 50%, for language family classification 20%, and for the language identification task 3.33%.

4.1 Individual feature sets

The accuracy results for each feature set described in Section 3.4 for the in-domain evaluation scenario are presented in Table 4.1.

Feature Set	Binary	Families	NLI
Char. 3-grams	85.46	73.27	57.11
Token unigrams	86.20	60.73	29.90
Spelling	69.42	45.72	22.92
Grammar errors	65.08	32.23	5.95
FW	79.54	50.89	17.61
POS 3-grams	69.98	42.70	11.88
Sentence length	51.00	20.13	3.27
Social network	57.57	25.39	5.10
Subreddits	86.91	79.92	66.38

Table 4.1: In-domain accuracy, individual feature sets

Evidently, almost 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, 80% on the language family task and 66% 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, practically at chance level, even on the binary task; we therefore do not use it when we combine features below. 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 4.2 depicts the results obtained by combining character trigrams, tokens, and spelling features (Section 3.4.1). 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	90.63	76.07	64.63
Out-of-domain	81.25	60.03	35.27

Table 4.2: Results: content features

The content-independent features (Section 3.4.2), whose contribution is depicted in Table 4.3, indeed fare worse, but are seemingly more robust outside the domain of training.

	Binary	Families	NLI
In-domain	82.00	53.26	17.79
Out-of-domain	70.84	45.48	14.23

Table 4.3: Results: content-independent features

Table 4.4 shows the results obtained by combining the spelling features (Section 3.4.1) with the grammar features (Section 3.4.2). 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	72.89	49.06	22.93
Out-of-domain	70.53	39.35	12.50

Table 4.4: Results: grammar and spelling features

Table 4.5 shows the accuracy obtained by all the centrality features (Section 3.4.3), 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	57.57	25.39	5.10
Out-of-domain	56.42	25.04	4.54

Table 4.5: Results: centrality features

Finally, the contribution of the most popular subreddits feature is shown in Table 4.6. 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	86.91	79.92	66.38
Out-of-domain	84.16	81.11	62.42

Table 4.6: Results: most popular subreddits

Therefore, we report the results obtained with *all* features, with (Table 4.7) and without (Table 4.8) the reddit-specific most popular subreddit feature.

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 30 native languages, can be accomplished with very high accuracy, as high as almost 81% when evaluated in-domain, and 72% out-of-domain (Table 4.7). 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 five-way task of identifying the language family of the authors' L1 (Table 4.8).

	Binary	Families	NLI
In-domain	92.07	87.38	80.7
Out-of-domain	88.46	80.80	72.1

Table 4.7: Results: all features

	Binary	Families	NLI
In-domain	90.77	78.31	63.04
Out-of-domain	82.21	57.90	32.73

Table 4.8: Results: all features except subreddits

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. We evaluated the accuracy of identifying correctly English-speaking users. [1]

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.

4.4 Robustness across datasets

To further evaluate the robustness of our model, we experimented with a different dataset: the TOEFL corpus (Blanchard et al., 2013), consisting of essays authored by nonnative English speakers applying for undergraduate studies in the United States. The essays were written in response to three different prompts, and as the proficiency level of the authors is much lower than that of the Reddit authors (Rabinovich et al., 2018), we expect the signal of the authors' L1s to be more pronounced. On the other hand, in light of the completely different genre and domain, models trained on Reddit cannot be expected to perform too well on the TOEFL

[1] This experiment was performed on an earlier version of the dataset so results on the current dataset may slightly differ.

dataset.

We selected from the TOEFL dataset all the essays that were written by native speakers of languages that are also in the Reddit dataset: French, German, Italian, Spanish, and Turkish. This yielded 5000 essays, 1000 for each L1. Typically, the length of the essays in the TOEFL corpus is relatively short, and in particular, much shorter than the 100 sentences we used for evaluating the Reddit results. We therefore experimented with two scenarios for evaluation: one in which complete essays are used for testing, as is done in most approaches to NLI with the TOEFL dataset (Malmasi et al., 2017); and one in which we only evaluate on essays longer than 20 sentences, to make the results more comparable with the Reddit evaluation results reported above. In this latter configuration, after balancing the set, only 64 essays were retained for each L1. Furthermore, we also use two training scenarios, one in which we train on chunks of 100 sentences and one in which the chunk size for training is 20 sentences. In both cases, we trained on Europe ("in-domain") chunks from the Reddit dataset, after reducing it to contain only chunks written by authors from these five countries.

In all the experiments we used combinations of the features that could be applied to the TOEFL essays: function words, token n-grams, character n-grams, POS n-grams, sentence length, spell checker suggestions and grammar checking. The naïve baseline accuracy result for these experiments is obviously 20%.

The results are presented in table 4.9; they range between 36.5% and 41.5%. The best result, 41.5% accuracy, was obtained when shorter chunks were used for training, and longer essays were used for testing. The results far outperform the baseline, strongly indicating the robustness of our model. Still, they are not as good those achieved by testing on the Reddit dataset. This is unsurprising given the stark differences between the two datasets.

Training/Testing	All essays (5000)	Longer essays (320)
Chunks of 100 sentences	37.6	36.5
Chunks of 20 sentences	38.4	41.5

Table 4.9: TOEFL experiment results

Chapter 5

Analysis

We now set out to analyze some of the more interesting classification 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 5.1 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.

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).

	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 5.1: Centrality features: average values and standard deviation

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'. Spu-

rious occurrences of *'e'* were also very common among all nonnative users, especially authors whose L1 was French: *'governement'*, *'unemployement'*, *'explicitely'*. 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.

Chapter 6

Deep learning approaches

In recent years, deep learning became more and more dominant in the NLP world and has contributed greatly to the improvement of many NLP tasks. Different kinds of neural networks have been employed to achieve state-of-the-art results for several NLP tasks (Goldberg, 2017). In this section we investigate whether neural networks can improve the classification results for the NLI task, too.

6.1 Model

We used a standard recurrent neural network model, Long Short-Term Memory (LSTM) with an attention mechanism,[1] since this model is known to effectively learn structure from sequential data and has been shown to perform well on many NLP tasks. The model consists of the following blocks: an embedding layer with word vector size of 300 tokens, an LSTM layer with dropout of 60%, an attention mechanism as described in Bahdanau et al. (2015) with vector dimension of 300, and a feed forward classification layer with softmax activation. All layers, including the embeddings layer, are initialized randomly. The classification layer consists of 30 neurons, one for each language. We used the Adam optimizer (Kingma and Ba, 2015), a categorical cross-entropy loss function, a learning rate of 0.003, and batch size of 32. Figure 6-1 graphically depicts the model.

The model receives as input chunks of 1500 tokens, roughly comparable to the 100 sentence chunks that were used with the classifiers described above. Chunks with more than 1500 tokens

[1] We used Keras for the implementation.

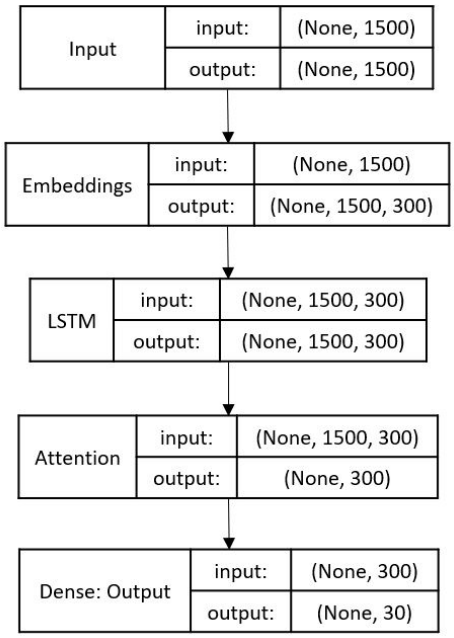

Figure 6-1: The LSTM model

were truncated, and chunks with fewer tokens were padded with zeroes. The chunks were preprocessed with the Keras tokenizer in order to convert the text to integers. This tokenizer receives as an argument a vocabulary size, k, and converts only the top-k most frequent words in the training chunks into integers, ignoring all other words. During tuning of the network parameters, we noticed that by decreasing the size of the vocabulary k we can improve the results. The original size of the vocabulary in our training data was approximately 100,000.

6.2 Results

We focus on the full (30-way) NLI classification task. The results, as a function of the vocabulary size k, are described in Table 6.1 A vocabulary size of only 5,000 word types yielded the best results both in-domain and out-of-domain. Since we feed to the neural network raw text chunks, without any feature engineering, these results should be compared to those obtained by all the features except subreddits with the classic model (Table 4.8). The LSTM results in a great improvement compared to the feature-based classifier: 73.63% accuracy for the in-domain NLI task (compared to 63.04% with the classic model), and 51.04% for the out-of-domain task (compared to 32.73% with the classic model). An interesting result was that

using a vocabulary size as small as 500 word types yielded decent results of 51.61% in-domain and 34.59% out-of-domain.

Vocabulary size	in-domain	out-of-domain
100000	43.6	29.2
50000	72.1	37.6
5000	73.6	51.0
500	51.6	34.6

Table 6.1: Accuracy results: LSTM

6.3 Variations

Combining spelling features Since the spelling features seemed to contribute greatly to the accuracy of the content features in the classic model, we thought it would be interesting to try to combine these features with the neural network model, in order to see if they can improve the results even more. We used the same model described above, but before the final classification layer, we added a concatenation layer that merged the output of the previous layer with a spelling features vector. The spelling vectors were extracted in the same way as in the classic model and used as additional inputs to the network. In addition, we added an extra fully-connected layer with 64 neurons, connected to the concatenation layer. We used a vocabulary size of 5,000 since this value maximized the accuracy with the basic LSTM model. The combined model is described in Figure 6-2.

Unfortunately, this augmented model did not improve the results. We therefore added to the combined model an auxiliary output, connected to the attention layer, identical to the main output layer. Thus, the weights of the main route are adjusted to improve the classification even before the merge with the spelling inputs (Figure 6-3). This model yielded a small improvement for the in-domain task (77.15% accuracy) but the results for the out-of-domain task were slightly lower (48.72%).

Model	in-domain	out-of-domain
LSTM with spelling	69.27	44.48
LSTM with spelling and aux output	77.15	48.72

Table 6.2: Accuracy results: LSTM with spelling features

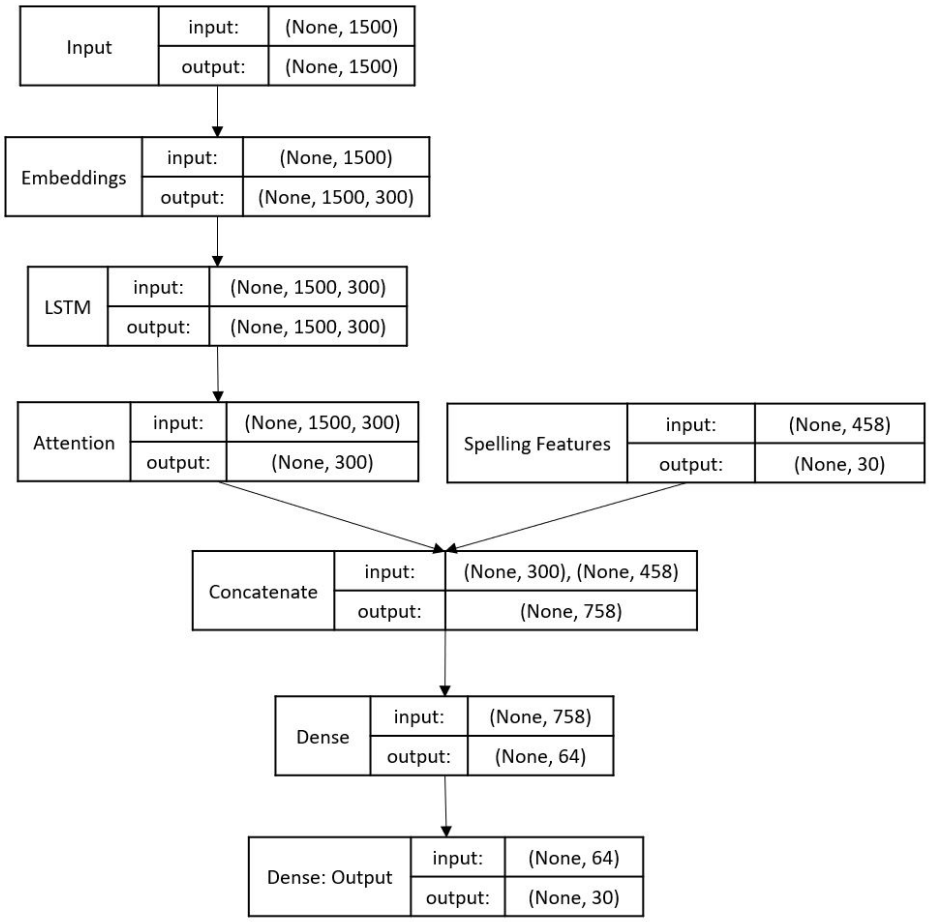

Figure 6-2: The LSTM model, augmented by spelling features

Trained word embeddings The LSTM results discussed above were obtained with randomly initialized word embeddings. We now describe experiments with trained embeddings, using a similar LSTM architecture implemented with the AllenNLP library (Gardner et al., 2017). We experimented with GloVE embeddings (Pennington et al., 2014), both learnable and fixed; and with the state-of-the-art ELMo embeddings (Peters et al., 2018), both off-the-shelf and ELMo embeddings trained on the Reddit corpus[2] As above, we experimented with two vocabulary sizes, 100,000 and 5,000. The accuracy results on the full (30-way) NLI task, in the in-domain scenario, are listed in Table 6.3.

Evidently, the results are disappointing. None of the more sophisticated models we tried was able to significantly outdo the basic LSTM model. A possible explanation for these results is that we did not have sufficient data for training complicated neural models, such as

[2]We are grateful to Nikita Haduong and Noah Smith for providing us with these pre-trained word embeddings.

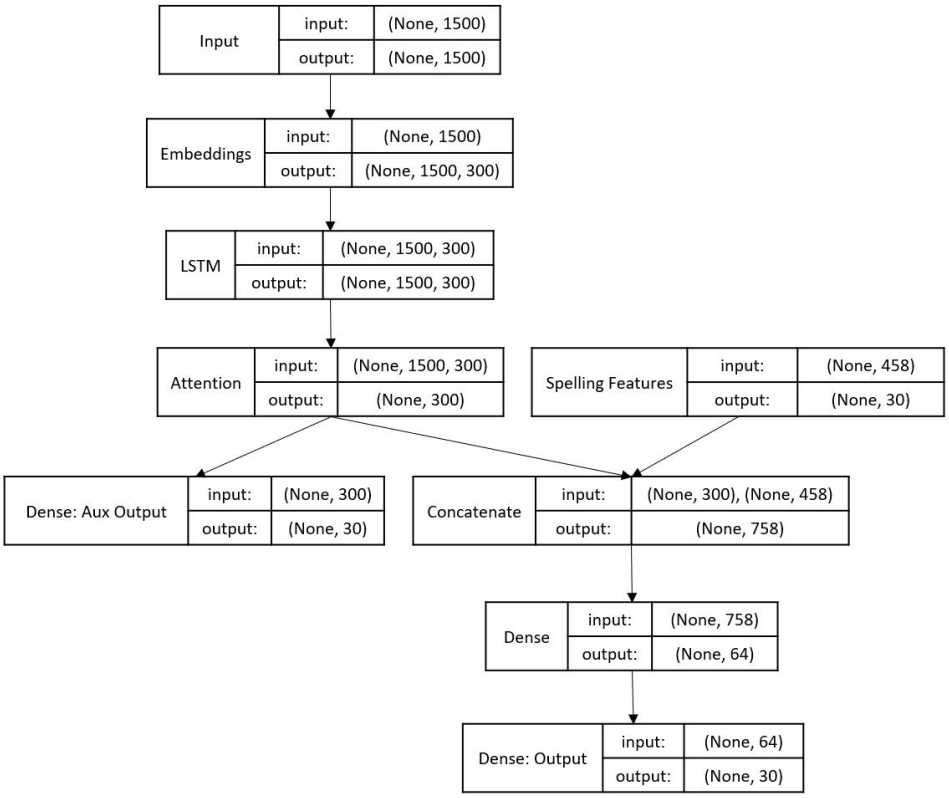

Figure 6-3: The LSTM model, augmented by spelling features and an auxilary output

Vocabulary	GloVE, trainable	GloVE, fixed	ELMO	ELMO, Reddit-trained
100k	30.13	29.91	12.01	20.11
5k	28.87	27.30	8.20	33.14

Table 6.3: In-domain NLI accuracy results, pre-trained word embeddings

ELMo embeddings. This may have caused the network to overfit to the training data and yield poor results on the test data. Also, we conclude that the lexicon of the reddit dataset may be very different from other datasets which were used for pre-training known embeddings such as GLoVE or ELMo. This is supported by the fact that using self-learned word embeddings yielded better results than both GLoVE and ELMo, and also by the fact that Reddit-pre-trained ELMo embeddings yielded better results than the default.

To alleviate some of these issues, we would like to downsample the data differently, focusing on more users per country rather than more sentences per user. Unfortunately, this will reduce the number of countries in the dataset. Another direction may be to find a configuration in which we use a larger portion of the out-of-domain data (where many chunks are unused),

e.g., using some of the out-of-domain subreddits also for training the model and keeping the rest of the out-of-domain subreddits for testing.

Chapter 7

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 more advanced language-model-based word embeddings (e.g., BERT (Devlin et al., 2018)) as well as with adversarial learning models for this task.

Bibliography

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. 2017. AllenNLP: A deep semantic natural language processing platform. ArXiv:1803.07640.

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

Yoav Goldberg. 2017. *Neural Network Methods for Natural Language Processing*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool.

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.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

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. *Transactions of the Association for Computational Linguistics*, 6:329–342.

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.