Native language binary classification

Gilad Eini and Liat Nativ

Abstract

Most of the world's population is bilingual. Moreover, even though the dominant language in the internet (particularly in social media) is English, there are evidences that most of the dynamic content is created by non-native English speakers. That is why the problem of distinguishing between native and non-native speakers is drawing attention. Potential applications of this task are: Teaching English more efficiently, identifying target audience based on native language etc. In this work we took an approach that is content independent - we are modelling text by the function words which occur in it. We then use machine learning techniques to distinguish between native and non-native English speakers, yielding solid results.

1 Introduction

Native language Identification (NLI) is the task of determining an author's native language based on his writing of a different language. NLI is a well-studied task introduced by Koppel (2005). This work is focused on the similar somewhat simplified task of Native language binary classification – based on a text written in English – decide if the writer is a native (English) speaker, or not. One popular approach for classification is Bag of Words (BOW). In BOW the text is represented by a feature vector of dimension N (N is the size of the vocabulary or a subset of the vocabulary- i.e. N most frequent words). Each entry in the feature vector is the count of the corresponding word form the vocabulary occurring in the text. Usually stop words are excluded from the feature vector. This approach (as will be shown in the results section of this document) is very effective for this task of classification.
However, this approach suffers from a main disadvantage which is content dependency. For instance, if a classification model is trained on specific domain corpus (e.g. Sports, Politics, traveling etc) this model will consider words from the specific domain with much higher importance compared to words out of that domain. This presents an issue in generalization across domains. Also there is a time relevance issue – domain dominant words can change significantly over time (this is particularly noticeable in domains such as politics and sports – where things changes rapidly). This could lead to poor classification results that might force retraining of the model for the new domain or the new period.

In order to overcome the content dependency issue we took a content independent approach by observing the use of function words (which do not carry content) for native and non-native English speakers. Because function words are not domain nor period dependent - this approach is robust to the issues described above and yields firm results for binary classification.

We use 2 different data sets for native and non-native English speakers respectively. Native language data set was extracted from Reddit – a popular American discussion website. We had 6 different countries of origin for native English speakers (USA, UK, Ireland, New-Zealand, Canada and Australia) for each we sampled an equal portion. Non-native English speakers dataset was taken from TOEFL (Test Of English as a Foreign Language) – a collection of assays written by non-native English speakers as a university entrance test. The country of origin of the writer was not specified. We used function words dictionary from the academic resources of Sequence Publishing as a base and extended it manually during our training process.

We used several well-known classifiers (SVM, Decision Tree and Naïve Bayes) for this task and compared the results of the function words approach vs. BOW approach. In addition, as far as we are aware of, most of NLI tasks does not consider more than one native English. To examine the effect of the different "Englishes" on the classification results we trained our model only on US-English vs. non-native English speakers.

2 Related Work

The task of native language identification (NLI) has got a fair amount of attention (Koppel, Schler, and Zigdon 2005, Wong and Dras 2011, Wong, Dras, and Johnson 2011, Swanson and Charniak 2012 Tetreault, Blanchard, et al. 2012, Jarvis & Paquot 2015 and many more).

2 Shared tasks were held - 2013 and 2017 in which dozens of teams were competing. In this work our focus was classification and not identification of native language – i.e. determine if a text was written by an English native speaker or not (binary classification).

McNamara & Crossley (2011) focused on identifying shared lexical features of non-native English speakers and found that some of these shared features are style related.

The approach taken in this work - of observing the use of function words for classification - was used before for the similar task of identifying translations source language (Koppel & Ordan 2011). This approach loosens the dependency upon content and focuses on stylistic characteristics. As shown (Koppel & Ordan 2011) some function words are over-represented and under-represented according to the source language, creating a solid base for classification.

3 Dataset

We used the Reddit dataset released by Rabinovich et al. (2018) as our native English speaker corpus. Reddit is a popular online community consisting of thousands of forums in a wide range of topics. The dataset includes Reddit posts 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, (though not perfect). Native English speaker countries used in this work consist of: US, UK, Australia, New Zealand, Ireland and Canada. For our non-native data set we used TOEFL dataset. The TOEFL test is used internationally as a measure of academic English proficiency, among other purposes, to inform admissions decisions for students seeking to study at institutions of higher learning where English is the language of instruction. The dataset consists of essays written by authors of different native languages on 8 topics sampled as evenly as possible.

4 Methods

4.1 Preprocessing

As a first step we converted datasets to lower case to avoid redundancy. Since our methodology focuses on function words we needed sentences that are in sufficient length. We sampled sentence containing at least 15, 20 and 45 tokens in them. The 45 tokens lower bounded sentences gave best scores. As TOEFL was the smaller dataset, we fixed the sample size of the larger dataset of Reddit accordingly. For the non-native English speakers, we had 11044 sentences for the 45 word sentences (across the entire dataset). For the native English speakers, we had 6 countries of origin to sample form. We randomly sampled an even portion of 1840 45-word sentences from each which summed up to 11040 sentences.

4.2 Feature vectors construction

4.2.1 Function words

For the sake of content independent classification, we relied on function words – words that are in general not related to content and considered as stylistic property.

We downloaded a dictionary from the academic resources at Sequence Publishing

Our dictionary consists of 311 function words. Our hypothesis is that native speakers use function ways in a way that is different than non-natives – and this will result in good classification results. For each sample we created a 311-dimension feature vector, where the kth entry in the vector is the number of occurrences of the kth function word from our dictionary in the given sample (sentence).

Looking at the top 10 frequent function words we found that native and non-native speakers share the same frequent words, almost in the same order. Non-natives however use these words more frequently, roughly 15% more than natives. Examining the difference manner natives and non-natives use function words, we present some of these differences in table 1.

Function words			
word	Count Native	Count Non-native	Difference
fewer	15	480	93.94%
against	513	49	82.56%
here	613	106	70.51%
therefore	62	289	64.67%
example	288	1279	63.24%
above	81	316	59.19%
must	125	433	55.2%
Into	870	316	46.71%

Table 1 function word usage difference examples

4.2.2 BOW

To examine the effectiveness of the function words approach, we confronted it with the Bag of Words (BOW) approach, that is known to yield excellent results, and is strongly depend upon content. We took the 230 most frequent words from the 2 datasets, leaving function words and other special characters out and removing duplicates – summing up to 343. We constructed a 343-dimension feature vector-where the kth dimension in the vector is the number of occurrences of the kth word from the 343 most frequent words we collected -in the given sample (sentence).

4.3 Classifiers

To perform the described classification task, we used 3 different classifiers and evaluate the results on all 3. The classifiers we used are SVM, Decision tree and Naïve Bayes.
For SVM, we used sklearn.svm as our main classifier. We used an RBF kernel, cost 1, and to improve run time we increased the cache size to 7000. For Naïve base we used sklearn naive set with the default parameters. For Decision Tree we used sklearn tree as the third classifier set with the default parameters. For all 3 we divided our data to 80% training set and 20% test set. We also ran 5 and 3-fold cross validation but that did not influence the results (most likely since we had enough data). In addition, we tested other sizes of training set, see 4.4.3

4.4 Other Settings

To get optimal results we changed some of the parameters during the experiments we made and we elaborate on the details here:

4.4.1 Sentence length

To avoid extremely sparse representation of function words in a sample (sentence) – it should contain enough tokens. On the other hand, selecting only very long sentences may reduce the amount of data noticeably. We tested our results on sentence length lower bound of 10, 25 and 45. Best results are achieved with the longest sentences. Results of this experiment are shown in the results section below, table 4.

4.4.2 BOW feature vector size

To compare the domain-independent approach with a domain-dependent one we used BOW of most frequent words in the vocabulary. To get best results for this method, we examined the results on a list of the 230, 500, 1000, 2000 and 3000 top frequent word for each one of the classes – resulting in a feature vector of dimension 343, 728, 1445, 2848 and 4725 respectively. Results of this experiment are shown in the results section below, table 5.

4.4.3 Training set size

To test the strength of the separation of the proposed method, we were interested in finding the minimal amount of data that was needed to train the model and get reasonable results. For that we set a fixed size training set of 2000 samples, and trained our model on a training set ranging from 2500 to ~20,000 samples (increasing the size by 2000 samples each time). We compare the results against BOW with same training set size.
Results of this experiment are shown in the results section below, table 6.

4.4.4 Different "dialects" of English

For our native English we sampled authors from 6 different English speaking countries: US, UK, Australia, New-Zealand, Ireland and Canada. As most of NLP tools and knowledge of English is focused on US English, we were interested in testing the influence of this mixture of "Englishes" on the results. For that we compared the original setting of 6 types of native English to solely US English as our native data set.

5 Results

For function word based classification we achieved best results with SVM using at least 45 tokens sentences. BOW best results were achieved with Naïve Bayes using top 343 most frequent words. We divided our data 80-20%. We measured the precision, recall and accuracy for each classifier on both methods. Table 2 shows the accuracy and f-score for function word classification. Table 3 shows the results of the BOW classification.

	Function words			
	Non-native f score	Native f score	weighted f score	Acc.
SVM	0.819	0.823	0.821	82.18 %
Decision tree	0.695	0.697	0.696	69.66 %
Naive Bayes	0.799	0.795	0.797	79.76 %

Table 2 – best results for function based classification

	Bow			
	Non-native f score	Native f score	weighted f score	Acc.
SVM	0.931	0.935	0.933	93.38%
Decision tree	0.876	0.874	0.875	87.52%
Naive Bayes	0.934	0.934	0.934	93.45%

Table 3 – BOW classification results.

.

As described in 4.4.1 above we tested the effect of sentence length on the results, shown in table 4 below (accuracy).

As described in 4.4.2 above we tested the effect of the amount of words considered for BOW classification. Accuracy results are shown in table 5 below. We used 343 most frequent words since it got the best results for SVM – which achieved the best results for function words based classification.

As described in 4.4.3 above we tested the effect of the training set size on the results. Table 6 shows 0ur experiment accuracy results. As can be seen in the table- results improve with the extension of the dataset, but fair results are achieved even with a training set of only 2500 samples.

As described in 4.4.4 above we tested the effect of using only US English as native class. Table 7 shows the accuracy results of this experiment. Results show improvement when narrowing the native English to US-English only.

	Function words		
Length	10	25	45
SVM	71.51%	76.52%	82.18%
Decision tree	62.23%	66.53%	69.66%
Naive Bayes	72.1%	75.66%	79.76%
	BOW		
Length	10	25	45
SVM	83.9%	89.47%	93.38%
Decision tree	81.16%	83.72%	87.52%
Naive Bayes	86.43%	90.46%	93.45%

Table 4 – results on different sentence length

Top X words from class	Feature vector size after removing duplicates	SVM	Decision Tree	NB
230	343	93.39	87.53	93.46
500	728	93.43	88.66	94.7
1000	1445	92.87	88.54	96.4
2000	2848	90.9	88.86	97.17
3000	4275	89.13	89.09	97.44

Table 5 – BOW different feature vector size

	Function words			BOW		
	SVM	Decision tree	Naive Bayes	SVM	Decision tree	Naive Bayes
2500	78.9	66.4	79	88	85.8	92.5
4500	79.9	67.3	79.5	90	86.8	93.1
6500	80.1	68.2	79.5	91.1	85.9	92.7
8500	80.7	66.7	79.7	91.7	86.9	92.7
10500	80.9	67.6	80.2	91.7	86.7	93
12500	81.1	68.5	80.6	92.3	86.6	92.9
14500	81.5	67.4	80.3	92.5	86.4	93
16500	81.7	67.7	80.4	92.6	86.2	93
18500	81.9	69.3	80.5	92.9	86.8	93.1
20084	82	69.3	80.5	93	86.8	93.1

Table 6 – Training set sizes

	Function words	BOW
SVM	82.5	93.7
Decision tree	70.2	88.72
Naive Bayes	80.08	93.88

Table 7 – accuracy classification results for US-English only.

6 Conclusions and Future Work

In this work we addressed the well-studied task of native language binary classification. We were interested in focusing on non-content dependent classification which is more robust and will presumably generalize well across domains and over time. We used function words, which, by definition, does not hold content as our feature. We got good results (more than 82% accuracy) on the function word based classification (based on 311 function words). We compared our results to the content-based BOW approach (based on 343 most frequent words in the corpus) which got excellent classification results (more than 93% accuracy and up to 97% when increasing feature vector size), but is less generalizable and will presumably perform poorly when dealing with different domains.

Future work can include testing this model on different dataset. We plan to take this task to a similar more challenging task of multi-class classification (also known as NLI – Native Language Identification) using deep learning techniques (most likely RNN)

References

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.Jarvis, Scott ; Paquot, Magali. Native language identification. In: Granger S. ; Gilquin G. ; Meunier F., Cambridge Handbook of Learner Corpus Research, Cambridge University Press : Cambridge 2015.

Aniket Kittur, Ed H. Chi and Bongwon Suh. Crowdsourcing User Studies With Mechanical Turk: Amazon.com's Mechanical Turk.

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

Wong, S.-M. J., Dras, M., & Johnson, M. (2011, December). Topic modeling for native language identification. In Proceedings of the Australasian Language Technology Association Workshop 2011 (pp. 115–124). Stroudsburg, PA: Association for Computational Linguistics.

Swanson, B., & Charniak, E. (2012, July). Native language detection with tree substitution grammars. In Proceedings of the 50th annual meeting of the Association for Computational Linguistics (Vol. 2; pp. 193–197). Stroudsburg, PA: Association for Computational Linguistics

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. San Rafael, CA: Morgan Claypool

Danielle S. McNamara, Scott A. Crossley, (2011) Shared features of L2 writing: Intergroup homogeneity and text classification

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.