World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:17, No:1, 2023

Native Language Identification with Cross-Corpus Evaluation Using Social Media Data: 'Reddit'

Yasmeen Bassas, Sandra Kuebler, Allen Riddell

Abstract—Native Language Identification is one of the growing subfields in Natural Language Processing (NLP). The task of Native Language Identification (NLI) is mainly concerned with predicting the native language of an author's writing in a second language. In this paper, we investigate the performance of two types of features; content-based features vs. content independent features when they are evaluated on a different corpus (using social media data "Reddit"). In this NLI task, the predefined models are trained on one corpus (TOEFL) and then the trained models are evaluated on a different data using an external corpus (Reddit). Three classifiers are used in this task; the baseline, linear SVM, and Logistic Regression. Results show that content-based features are more accurate and robust than content independent ones when tested within corpus and across corpus.

Keywords—NLI, NLP, content-based features, content independent features, social media corpus, ML.

I. Introduction

NATIVE Language Identification (NLI) task is concerned with predicting the native language of texts written by learners of a second language (L2). NLI relies on the assumption that speakers of the same native language display certain linguistic patterns in their L2 texts which can be used as traces in NLI to predict their L1. Work on NLI has exploited various types of these linguistic features such as function words, character n-grams, POS n-grams, syntactic structure, and spelling mistakes [1], [2].

Previous work (e.g., [3], [4]) has argued for content independent features as opposed to content-based features because they are less biased to the prompt and domain of the data. Content-based features consist of word n-grams while content independent features are non-lexical (such as POS tags or function words) and are thus less dependent on text vocabulary. Content-based features are considered less desirable because they cause topic bias since they depend on the topic of the text [3] and may thus not be useful for texts on different topics (e.g., different prompts in TOEFL).

In the current paper, we will investigate this argument in a cross-corpus setting, using social media data from Reddit[1]. The contribution of this paper is to examine which types of features will be more accurate and informative when tested across corpus using Reddit data. Using such a corpus makes the task harder since it is a totally different domain than learner corpora. Two types of linguistic features, content-based features and content independent features, are investigated in order to find out which ones work best when models are tested across corpus. Our hypothesis is that content-independent features are supposed to perform better when tested across corpus while content-dependent features are supposed to be more informative and accurate when tested within corpus. This is because content independent features are less biased to texts contents and topics. To our knowledge, this is the first attempt in NLI that examines the performance of these two types of features on such a different corpus like Reddit; a data of a highly-advanced non-native speakers of English.

The rest of the paper is organized as follows: Section II describes related work, Section III the data sets, Section IV explains our methodology, Section V shows our results and a discussion, and Section VI presents the conclusion and future work.

Y. Bassas and S. Kuebler are with the Department of Linguistics, Indiana University, Bloomington, IN, 47405, USA (e-mail: ybassas@iu.edu, skuebler@indiana.edu).

A. Riddell is with the Department of Information and Library Science, Indiana University, Bloomington, IN, 47405, USA (e-mail: riddella@indiana.edu).

[1]We use the data set collected by [5].

II. Related Work

In textual analysis tasks like NLI, various types of linguistic features have been exploited. Certain linguistic features such as words, lemmas, tokens, and characters n-grams have shown to be effective (e.g., [2], [6], [7]). However, these types of linguistic features can be problematic since they are content/domain dependent. By using such features, topic bias can occur specially when prompts or topics are not equally distributed across texts. This will cause the classifier to be indirectly trained on topics. Not to mention that these features will be also corpus-specific. Therefore, other research in NLI (e.g., [3], [8], [9]) reported the importance and usefulness of using features that are content independent to avoid these issues. Reference [4], for example, employed language independent features such as function words, POS n-grams, and mixed POS and function words. Results showed that these features were very effective in discriminating L1 groups. Similarly, [10] conducted a large-scale cross-corpus evaluation using language independent features (function words, POS n-grams, and CFG production rules). Even though within corpus results were better than across corpus results, the types of features used still showed discriminative power in distinguishing L1 groups. The drop in the performance of cross corpus evaluation suggested that these features were still corpus/genre dependent. Lately, [11] introduced an NLI task using a social media data namely Reddit. They used content-dependent features (token n-grams, character n-grams, and spelling and grammar errors), and content-independent features (function words, POS n-grams, sentence length, and social network features). Results showed that content dependent features were more accurate within domain while

content independent features were more robust when tested out of domain.

As seen above, some NLI studies argued for content independent features, this paper investigates the performance of content-based features and content independent ones in cross-corpus setting using Reddit data in order to find out which of these features work best in such a setting.

III. Data

Two corpora are used in this paper. The first one is TOEFL11 data [12] which is comprised of 12,100 English essays written by speakers whose native languages are Arabic, Chinese, French, German, Italian, Hindi, Japanese, Korean, Spanish, Telugu, and Turkish (11 languages). Data are divided to 90% as training set and 10% as test set. Data are POS tagged using Stanford POS tagger. In this paper, we run the experiments using the tokenized form of the data.

The second data set is collected from Reddit. We use the Reddit corpus released by [5]. These data consist of posts created by users who self-report a country affiliation using a Reddit feature called "flair". In this case, users use their flair to display a national flag. These flairs, following [5], are viewed as accurate information for the native language of the users (authors). As shown by [5], the English used by non-native speakers on Reddit is highly advanced which makes NLI task more challenging. Each post is associated with a unique user ID. The data contain around 50 countries most of them are European countries. The data are collected from European subreddits (such as r/europe, r/AskEurope, r/EuropeanFederalists) and then extended to non-European subreddits (such as r/AskReddit, r/IAmA, r/funny) by mining posts of the users who previously declared their flairs in European subreddits. For our test set, we extract 1,800 posts (300 posts per language) based on 13 subreddits found in the non-European subreddits set. We focus on 5 native languages that exist in TOEFL data (French, German, Italian, Spanish, and Turkish). We POS tagged the data using the Stanford POS tagger. Below we show examples from TOEFL and Reddit data.

TOEFL excerpts:

- It is obvious that advertisements do not provide a complete information to the customers.
- It could be a problem if the advertisement say lies during its presentations or present inexact information that could guide the consumers to a invalid choice.
- First of all in past times people especially youth has much more spare time compared to present times youth.
- So young people meet more people, make more friends, have more fun, and have less problems to worry about.

Reddit excerpts:

- That's a really great question actually. I think it'd have to be a game I played against a friend in Rome 2,
- Yes, they got bought by a very weird/funny/creepy/out of his mind guy called Ferrero, who seems to be on cocaine or LSD the whole time.
- you will need to point it at your video player inside of your SVP folder
- Are you even paying attention to the context of these comments or are you just automatically taking the side of everyone with a Galatasaray flair?

IV. Methodology

A supervised multi-class classification method is employed using the following three classifiers: Linear Support Vector Machine, Logistic Regression and a simplified classifier used as a baseline [13].

The number of iterations chosen for training algorithm for Logistic Regression classifier is fixed to 100 while it is fixed to 1000 for linear SVM. We used the default setting. We used features that are commonly used and reported to be informative and accurate in NLI literature (e.g., [4], [6]). Feature weights consist of TF-IDF weights are utilized.

For content-based features, we use word n-grams ranging from 1 to 4, character n-grams from 2 to 11, and a combination of word unigrams/bigrams and character n-grams from 4 to 11. For content independent features, we use POS n-grams ranging from 1 to 8, a combination of POS n-grams ranging from 1 to 4, function words, and a combination of POS n-grams and function words.

The baseline is the majority class. Since the data set is balanced, we randomly chose one class.

V. Results and Discussion

We train our models within domain using TOEFL data and then we evaluate these models on an external corpus using Reddit data. Data are split into 90% train and 10% test.

A. Within Corpus Experiment Using TOEFL

In this experiment, we train and test our models within domain, i.e., using TOEFL data. We report two settings: one with only five L1s that are found in Reddit data, and one with all L1s found in the TOEFL data. The first setting allows a direct comparison to the across corpus experiment with Reddit data as test data. The second settings allow a comparison to prior work. We focus our discussion mainly on the first setting.

The results using the content-based features are shown in Table I, the results using the content independent features are shown in Table II.

The results based on content-based features (Table I) show that all settings perform significantly higher than the baseline. We also see that the SVM generally outperforms the logistic regression model. The SVM performs best with word unigrams, bigrams and character 5-grams while logistic regression reaches the highest performance using all word and character n-grams. It is worth noting that the results are rather unstable, and using all n-grams within a specific range does not always result in the best performance. The best setting for the SVM is based on around 900 000 features, as compared to around 23.5 million for all n-grams. We experimented with more combinations, but only report the the most interesting results.

When comparing the subset of 5 languages with the full set of languages, we see similar results, with the 5 languages reaching slightly higher accuracies. This is to be expected since the model has to choose between fewer classes.

When we look at the content independent features (Table II), it is immediately obvious that these results are considerably lower than the results with content-based features, even though

TABLE I
WITHIN CORPUS ACCURACIES OF CONTENT-BASED FEATURES FOR THE 5 REDDIT LANGUAGES AND FOR ALL LANGUAGES

		5 languages		TOEFL lang.		
Setting	Type	LR	SVM	LR	SVM	# Features
Baseline		20.00		9.09		
Words only	1+2	80.20	84.80	78.72	81.72	675 194
	3+4	55.80	64.80	53.72	62.54	4 536 405
Words & char. seq.	W2+C11	76.20	83.20	74.00	80.54	6 821 979
	W1+2+C5	81.20	**86.80**	80.18	**83.45**	911 470
	W1+2+C11	79.20	85.20	78.36	82.54	6 878 074
	W1+2+C_all	**82.20**	85.20	80.36	83.00	18 947 945
	W_all+C_all	**82.20**	84.80	**80.54**	83.00	23 484 350

TABLE II
WITHIN CORPUS ACCURACIES OF CONTENT INDEPENDENT FEATURES FOR THE 5 REDDIT LANGUAGES AND FOR ALL LANGUAGES

		5 languages		TOEFL lang.		
Setting	Type	LR	SVM	LR	SVM	# Features
Baseline		20.00		9.09		
POS only	3+4	53.00	52.60	54.00	53.00	129 863
	1+2+3	51.60	49.80	53.45	51.54	16 880
	1+2+3+4	53.00	51.40	54.54	52.63	130 924
Function words		39.20	40.00	41.81	42.09	367
POS & function words	1+2+3+fw	59.20	55.40	**60.18**	55.63	17 247
	1+2+4+fw	57.20	**58.60**	58.54	**58.27**	115 472
	1+2+3+4+fw	**59.40**	57.40	60.00	58.09	131 291

they outperform the baseline. Overall, logistic regression outperforms the SVM. The only exception is for function words, but this setting reaches the lowest results. The SVM prefers POS uni-, bi-, and 4-grams in combination with function words while the logistic regression performs best with POS uni-, bi-, and tri-grams in combination with function words.

The fact that the results show a substantial drop in performance in content independent features compared to content-based ones is not surprising since this experiment is samples from the same corpus, and thus the same topics, for training and test data. This is in line with findings by [6] and [7].

B. Cross-Corpus Experiment Using Reddit

Our second experiment tests the effectiveness of our features when evaluated across corpus, on Reddit data. Since the corpus texts are from different prompts and genres, this approach allows us to tests the generalizability of our features on a domain different from that of the training data. Table III shows the results of using content-based features, and Table IV shows the results of using content independent features.

The results of the content-based features (Table III) show the expected drop in accuracy. Results here are often below the baseline. Logistic regression only manages to surpass the baseline by 1.3% absolute with the combination of word bigrams and character 11-grams. The SVM performs better, reaching its highest accuracy of 25.86% when using word uni- and bigrams along with character 11-grams. It is also striking that while in the within-corpus setting, adding character n-grams resulted in an improvement of around 2% for the SVM, in the cross-corpus setting, the gain is minimal, from 25.40% to 25.86%.

Now, if the hypothesis that content independent features work better across corpus is true, we would expect this setting to be more robust. However, the results of this setting (Table IV) show an even more dramatic drop in accuracies. All results are considerably lower than the baseline. Especially the logistic regression, which performed well on the within-corpus setting using only content independent features, reaches the highest accuracy of less than 1%. The SVM is somewhat more stable, reaching its highest accuracy of 11.13% when using POS tri- and 4-grams.

The comparison of the results of the in-domain

TABLE III
CROSS-CORPUS RESULTS (ACCURACY) FOR CONTENT-BASED FEATURES

Type	LR	SVM
Baseline	20.00	
Words only		
1+2	18.66	25.40
3+4	19.53	20.20
Word and character n-grams		
W2+C11	**21.33**	24.73
W1+2+C5	17.20	23.93
W1+2+C11	19.06	**25.86**
W1+2+C_all	15.73	22.73
W_all+C_all	15.60	22.80

TABLE IV
CROSS-CORPUS RESULTS (ACCURACY) FOR CONTENT INDEPENDENT FEATURES

Type	LR	SVM
Baseline	20.00	
POS only		
3+4	0.82	**11.13**
1+2+3	**0.83**	10.46
1+2+3+4	0.80	0.96
Function words	0.42	0.34
POS and function words		
1+2+3+fw	0.53	0.72
1+2+4+fw	0.52	0.62
1+2+3+4+fw	0.55	0.75

TABLE V
WITHIN CORPUS RESULTS: P, R, F OF CONTENT-BASED FEATURES FOR 5 LANGUAGES (WORD 1,2 AND CHAR 5 BY SVM)

Languages	Precision	Recall	F1
German	90.48	95.00	92.68
Italian	90.00	90.00	90.00
Turkish	93.26	83.00	87.83
French	86.41	89.00	87.68
Spanish	95.06	77.00	85.08

TABLE VI
WITHIN CORPUS RESULTS: P, R, F OF CONTENT INDEPENDENT FEATURES FOR 5 LANGUAGES (POS 1,2,4 AND FW BY SVM)

Languages	Precision	Recall	F1
German	79.07	68.00	73.12
Italian	73.91	68.00	70.83
Turkish	73.24	52.00	60.82
Spanish	65.85	54.00	59.34
French	67.11	51.00	57.95

TABLE VII
CROSS-CORPUS RESULTS: P, R, F OF CONTENT-BASED FEATURES FOR 5 LANGUAGES (WORD 1,2 AND CHAR 5 BY SVM)

Languages	Precision	Recall	F1
German	75.96	26.33	39.11
Italian	33.89	40.33	36.83
French	50.39	21.33	29.98
Spanish	46.43	17.33	25.24
Turkish	58.11	14.33	22.99

TABLE VIII
CROSS-CORPUS RESULTS: P, R, F OF CONTENT INDEPENDENT FEATURES FOR 5 LANGUAGES (POS 1,2,4 AND FW)

Languages	Precision	Recall	F1
Spanish	24.37	9.67	13.84
German	32.08	5.67	9.63
French	35.71	5.00	8.77
Turkish	20.51	5.33	8.47
Italian	16.00	5.33	8.00

and out-of-domain experiments show very clearly that content-based features are more accurate and informative than content independent features when the models are evaluated within corpus or across corpus. Even though a drop in performance is expected due to the significant prompts/genres differences between two corpora, content-based features are still more robust than content independent ones. Although content independent features are believed to be less biased specially in settings across corpora [10], [11], in our scenario, results show that POS features do not generalize across corpus and are thus equally biased. This means that even though the content-based features are biased, they still provide better information across corpus.

C. Results per Language

We also look at the results per language since those show trends that differ from the general trend: we report results for both experiments for the five L1s in the Reddit data. We report results of the best model of both types of features. The results of content-based features are shown in Table V and Table VII.

The results of content independent features are shown in Tables VI and VIII.

A comparison of the content-based features results on both data sets show that using these features provides stable performance whether within corpus or across corpus. A comparison of the best performing languages in Tables V and VII shows a similar order except for Turkish (which is possibly harder to classify in cross-corpus setting). However, looking at content independent features results indicates that they are unstable and less predictable across corpus. This can be clearly shown in Tables VI and VIII where we notice a considerable variation in the order of the best languages between within corpus and cross-corpus experiment. We can conclude that the within corpus performance of content-based features is a good predictor of their performance across corpus; for content independent features, this is not the case.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have investigated how well content-based features vs. content independent features perform when evaluated across corpus. Results show that content-based features are more accurate when tested within corpus and across corpus. Content independent features prove to be less informative and less predictive when tested within corpus or on a different corpus, and their within-corpus performance is not predictable of their performance across corpus.

We are planning to investigate more types of features, such as spelling and grammar errors. Additionally, we need to compare our results to a neural architecture. Most importantly, we need to broaden the spectrum and look at L2s other than English, to determine if the same regularities hold.

REFERENCES

[1] M. Koppel, J. Schler, and K. Zigdon, "Automatically determining an anonymous author's native language," in *International Conference on Intelligence and Security Informatics*, ser. Lecture Notes in Computer Science, vol. 3495, 2005.

[2] S.-M. J. Wong and M. Dras, "Contrastive analysis and native language identification," in *Proceedings of the Australasian Language Technology Association Workshop*, Sydney, Australia, Dec. 2009, pp. 53–61. [Online]. Available: https://www.aclweb.org/anthology/U09-1008

[3] S. Malmasi and M. Dras, "Arabic native language identification," in *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 180–186. [Online]. Available: https://www.aclweb.org/anthology/W14-3625

[4] S. Malmasi, M. Dras, and I. Temnikova, "Norwegian native language identification," in *Proceedings of the International Conference Recent Advances in Natural Language Processing*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, Sep. 2015, pp. 404–412. [Online]. Available: https://www.aclweb.org/anthology/R15-1053

[5] E. Rabinovich, Y. Tsvetkov, and S. Wintner, "Native language cognate effects on second language lexical choice," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 329–342, 2018. [Online]. Available: https://www.aclweb.org/anthology/Q18-1024

[6] B. G. Gebre, M. Zampieri, P. Wittenburg, and T. Heskes, "Improving native language identification with TF-IDF weighting," in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, Atlanta, GA, Jun. 2013, pp. 216–223. [Online]. Available: https://www.aclweb.org/anthology/W13-1728

[7] I. Markov, L. Chen, C. Strapparava, and G. Sidorov, "CIC-FBK approach to native language identification," in *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 374–381. [Online]. Available: https://www.aclweb.org/anthology/W17-5042

[8] S. Malmasi and M. Dras, "Chinese native language identification," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*. Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014, pp. 95–99. [Online]. Available: https://www.aclweb.org/anthology/E14-4019

[9] S.-M. J. Wong and M. Dras, "Exploiting parse structures for native language identification," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, UK, Jul. 2011, pp. 1600–1610. [Online]. Available: https://www.aclweb.org/anthology/D11-1148

[10] S. Malmasi and M. Dras, "Large-scale native language identification with cross-corpus evaluation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 1403–1409. [Online]. Available: https://www.aclweb.org/anthology/N15-1160

[11] G. Goldin, E. Rabinovich, and S. Wintner, "Native language identification with user generated content," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct.-Nov. 2018, pp. 3591–3601. [Online]. Available: https://www.aclweb.org/anthology/D18-1395

[12] D. Blanchard, J. Tetreault, D. Higgins, A. Cahill, and M. Chodorow, "ETS Corpus of Non-Native Written English," Linguistic Data Consortium, LDC2014T06, 2013.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.