CIC-IPN@INLI2018: Indian Native Language Identification

Ilia Markov[1] and Grigori Sidorov[2]

[1] INRIA
Paris, France
ilia.markov@inria.fr
[2] Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC),
Mexico City, Mexico
sidorov@cic.ipn.mx

Abstract. In this paper, we describe the CIC-IPN submissions to the shared task on Indian Native Language Identification (INLI 2018). We use the Support Vector Machines algorithm trained on numerous feature types: word, character, part-of-speech tag, and punctuation mark n-grams, as well as character n-grams from misspelled words and emotion-based features. The features are weighted using log-entropy scheme. Our team achieved 41.8% accuracy on the test set 1 and 34.5% accuracy on the test set 2, ranking 3rd in the official INLI shared task scoring.

Keywords: Native Language Identification · Indian languages · social media · feature engineering · machine learning.

1 Introduction

The task of Native Language Identification (NLI) consists in identifying the native language of a person based on their text production in the second language. The underlying hypothesis is that the learner's native language (L1) influences their second language (L2) production as a result of the language transfer effect (native language interference) [14], which is thoroughly studied in the field of second language acquisition (SLA).

The possible applications of the task include marketing and security, as NLI is viewed a subtask of author profiling, as well as education, where the pedagogical material can be targetly tuned to native languages, for example, by taking into account the most common errors made by learners with a specific background and adapting the materials to tackle such errors in more detail.

Previous studies on identifying the native language from L2 writing – most of which approached the task from a machine-learning perspective – explored a wide range of L1 phenomena that appear in L2 production, i.e., lexical choices made by learners, grammatical patterns used, the influence of cognates and general etymology, spelling errors, punctuation, emotions, among others, and used corresponding features to capture these phenomena. Most of the NLI studies

focused on English as second language; however, NLI methods have been also examined on other L2s with promising results [8].

The interest in NLI led to the organization of several NLI competitions, including the first edition of the shared task on identifying the Indian languages [6], which was held in 2017 and attracted a large number of participating teams. The winning approach consisted in training the Support Vector Machines (SVM) classifier with SGD (Stochastic Gradient Descent) method on word n-gram and character n-gram features [5]. Other approaches included using several pre-processing steps (e.g., removing digits, emoji, stop words), classification algorithms (e.g., SVM, Logistic Regression, Naive Bayes), and features (e.g., non-English word counts, using adjectives and nouns as features, average sentence and word length, among others) [6].

In this paper, we present the CIC-IPN submissions to the INLI shared task 2018 [7]. We use the SVM algorithm trained on word n-grams, traditional (untyped) and typed character n-grams, part-of-speech (POS) tag n-grams, punctuation mark n-grams, character n-grams from misspelled words, and emotion-based features. In continuation we describe in detail the features used and the configuration of our runs.

2 Data

The training dataset released by the organizers consists of Facebook comments in the English language extracted from regional language newspapers. This dataset was also used in the 2017 edition of the INLI competition [6]. The dataset statistics in terms of the L1s covered, number (No.) of documents per L1, and the corresponding ratio are provided in Table 1. It can be seen that 1,233 training documents are nearly-optimally balanced across the represented L1s.

The submitted systems were evaluated on two test sets: the test set 1 (also used in the INLI 2017; 783 documents) and the test set 2 (the official test set of the INLI 2018; 1,185 documents).

Table 1. INLI training dataset statistics.

Language	**No. of documents**	**Ratio**
Malayalam (MA)	200	16.22%
Bengali (BE)	202	16.38%
Kannada (KA)	203	16.46%
Tamil (TA)	207	16.79%
Telugu (TE)	210	17.03%
Hindi (HI)	211	17.11%
Total	1,233	100%

3 Methodology

In this section, we give a description of the features incorporated in our runs and the configuration of our system: weighting scheme, frequency threshold, and machine-learning classifier.

3.1 Features

Word n-grams capture lexical choices of the learner in L2 production, and are considered one of the most indicative unique feature types for the task of NLI [4, 9]. Word n-gram features were also incorporated in the winning approach to the previous INLI shared task [5]. In runs 1 and 3, we use word unigrams and 2-grams, when in run 2 we use word 1–3-grams. We lowercase the word-based features and replace digits by a placeholder (e.g., 12345 → 0).

Untyped character n-grams are considered very indicative features for NLI and for other related tasks [3, 16]. In NLI, these feature are hypothesized to capture the phoneme transfer from the learner's L1 [18], among others L1 peculiarities. They were also incorporated into the winning approach to the INLI 2017 [5]. We use character n-grams with n = 2.

Typed character n-grams – character n-grams categorized into ten different categories – have been successfully applied to NLI [9]. We conducted an ablation study in order to identify the most indicative typed character n-gram categories. We found that the *middle-punctuation* and the *whole-word* categories did not contribute to the result, and therefore were discarded. We use typed character 4-grams; 3-grams are used for the *suffix* category.

POS tag n-grams capture morpho-syntactic aspects of the native language in NLI. They encode word order and grammatical properties of the native language, capturing the use or misuse of grammatical structures. POS tag n-grams have proved to be useful features for NLI, especially when combined with other feature types [4, 9]. We use POS tag 3-grams; obtaining the POS tags with the TreeTagger package [17].

Punctuation mark n-grams The impact of punctuation marks (PMs) on NLI was evaluated in [11]. The authors report that punctuation usage is a strong indicator of the author's L1. We use punctuation mark n-grams (n = 3).

Character n-grams from misspelled words were introduced by Chen et al. [2]. These features have been successfully used to tackle the NLI task in [9]. We extract 8,937 misspelled words (from the training dataset) using the PyEnchant package[3] and build character 4-grams from them.

[3] https://pypi.org/project/pyenchant/

Emotion polarity features Emotion-based features for NLI were proposed in [12]. We use emotion polarity (emoP) features similar to [12]: replace each word in the text with the information form the NRC emotion lexicon [13], e.g., *excellent*→ "0000101001".

3.2 Weighting scheme and threshold

We use log-entropy (le) weighting scheme that measures the importance of a feature across the entire corpus. le is considered one of the best weighting schemes for the NLI task [4, 2, 9]. In our experiments under 10-fold cross-validation, le outperformed other weighting schemes we examined (tf-idf, tf, and binary). Accuracy improvement over the second best-performing weighting scheme (tf-idf) was 3.2%–3.6% depending on the run.

Tuning the size of the feature set (selecting the optimal frequency threshold values) is an effective strategy for NLP tasks in general [10] and for NLI in particular [4, 9]. In all our runs, we include only the features that appear in two documents (min_df =2). In run 3, we additionally set frequency threshold value to 3 (include only the features that appear three times in the entire corpus).

3.3 Classifier

We use the linear SVM algorithm whose effectiveness has been proved by numerous studies on NLI [4, 9]. SVM was also the most popular algorithm in the 2017 edition of the INLI shared task [6]. SVM with OvR (one vs. the rest) multi-class strategy is used, as implemented in the scikit-learn package [15].

3.4 Evaluation

For the evaluation of our system, we conducted experiments under 10-fold cross-validation, measuring the results in terms of classification accuracy on the training corpus.

4 Results and Discussion

Table 2 summarizes the 3 runs submitted to the INLI shared task 2018 in terms of the features and the thresholds used. It also presents the 10-fold cross-validation results and the official results on the test sets 1 and 2 (accuracy, %).

As one can see from Table 2, there is a very high accuracy drop on the test data compared to the 10-fold cross-validation results. The drop is likely caused by the word and character n-gram features, which are known to capture not only peculiarities of the L1 but topic-related information as well [1]. The observed overfitting can be due to the size of the training dataset and/or presence of topic bias. Additional experiments are required to understand in more detail the reason for this performance.

Table 2. Summary of the three runs submitted by the CIC-IPN team and the obtained results (accuracy, %). Number of features for each run is also provided. 10FCV stands for 10-fold cross-validation.

Features	**Run1**	**Run2**	**Run3**
BoW	✓	✓	✓
Word 2-grams	✓	✓	✓
Word 3-grams		✓	
Character 2-grams	✓	✓	✓
Typed character n-grams (n = 3/4)	✓	✓	✓
POS 3-grams	✓	✓	✓
Character 4-grams from misspelled words	✓	✓	✓
Punctuation mark 3-grams	✓	✓	✓
Emotion polarity	✓	✓	✓
Min_df	2	2	2
Threshold	–	–	3
Number of features	75,667	92,789	51,886
10FCV accuracy	96.2%	96.0%	95.9%
Test set 1 accuracy	**41.8%**	41.3%	41.4%
Test set 2 (official) accuracy	34.1%	34.4%	**34.5%**

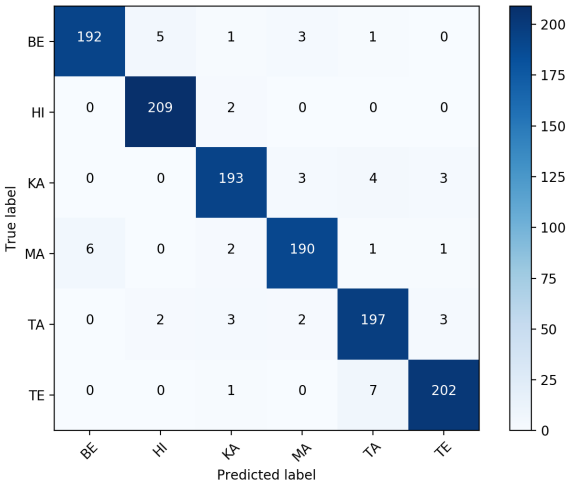

Fig. 1. Run 3: 10-fold cross-validation confusion matrix.

Run 3 showed the highest accuracy on the official test set due to a higher frequency threshold value. The confusion matrix for this run on the training data is shown in Figure 1; the class-wise accuracy results provided by the organizers on the test set 1 and 2 are presented in Tables 3 and 4, respectively. The highest 10-fold cross-validation result was achieved for the Hindi language, while on the both test sets this was the hardest language to identify.

Table 3. Run 3: class-wise accuracy results on the test set 1.

Class	Precision	Recall	F1
BE	55.0%	50.8%	52.8%
HI	44.2%	15.1%	22.6%
KA	33.8%	62.2%	43.8%
MA	39.3%	62.0%	48.1%
TA	32.7%	49.0%	39.2%
TE	42.1%	49.4%	45.5%
Overall accuracy		41.4%	

Table 4. Run 3: class-wise accuracy results on the test set 2.

Class	Precision	Recall	F1
BE	43.7%	28.5%	34.5%
HI	14.8%	13.8%	14.3%
KA	44.0%	44.0%	44.0%
MA	38.6%	36.5%	37.5%
TA	24.4%	50.7%	32.9%
TE	40.1%	30.8%	34.8%
Overall accuracy		34.5%	

5 Conclusions

We described the three runs that were submitted by the CIC-IPN team to the INLI shared task 2018. Our approach uses the SVM algorithm trained on word, character, POS tag, and punctuation mark n-grams, character n-grams from misspelled words, and emotion-based features. The features are weighted using log-entropy weighting scheme. Our team achieved 41.8% accuracy on the test set 1 (run 1) and 34.5% accuracy on the official test set 2 (run 3), placing our team 3rd (out of 12 participating teams) in the competition.

In future work, we will evaluate the performance of our system without word and character n-grams in order to investigate their impact on the accuracy drop suffered by the system when evaluated on the test sets. We will also focus on more abstract features that perform well in the situation where topic bias may occur.

References

1. Brooke, J., Hirst, G.: Native language detection with 'cheap' learner corpora. In: Proceedings of the Conference of Learner Corpus Research. pp. 37–47. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium (2011)
2. Chen, L., Strapparava, C., Nastase, V.: Improving native language identification by using spelling errors. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 542–546. ACL, Vancouver, Canada (2017)
3. Gómez-Adorno, H., Markov, I., Baptista, J., Sidorov, G., Pinto, D.: Discriminating between similar languages using a combination of typed and untyped character n-

grams and words. In: Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 137–145. ACL, Valencia, Spain (2017)
4. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. In: Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 111–118. ACL, Atlanta, GA, USA (2013)
5. Kosmajac, D., Keselj, V.: DalTeam@INLI-FIRE-2017: Native language identification using SVM with SGD training. In: Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation. CEUR, Bangalore, India (2017)
6. Kumar, A., Ganesh, B., Singh, S., Soman, P., Rosso, P.: Overview of the INLI PAN at FIRE-2017 track on Indian native language identification. In: Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation. vol. 2036, pp. 99–105. CEUR Workshop Proceedings, Bangalore, India (2017)
7. Kumar, A., Ganesh, B., Soman, P.: Overview of the INLI@FIRE-2018 track on Indian native language identification. In: Workshop proceedings of FIRE 2018. CEUR Workshop Proceedings, Gandhinagar, India (2018)
8. Malmasi, S., Dras, M.: Multilingual native language identification. Natural Language Engineering **23**(2), 163–215 (2017)
9. Markov, I., Chen, L., Strapparava, C., Sidorov, G.: CIC-FBK approach to native language identification. In: Proceedings of the 12th Workshop on Building Educational Applications Using NLP. pp. 374–381. ACL, Copenhagen, Denmark (2017)
10. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: The winning approach to cross-genre gender identification in Russian at RUSProfiling 2017. In: FIRE 2017 Working Notes. vol. 2036, pp. 20–24. CEUR-WS.org, Bangalore, India (2017)
11. Markov, I., Nastase, V., Strapparava, C.: Punctuation as native language interference. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3456–3466. The COLING 2018 Organizing Committee, Santa Fe, New Mexico, USA (2018)
12. Markov, I., Nastase, V., Strapparava, C., Sidorov, G.: The role of emotions in native language identification. In: Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis. ACL, Brussels, Belgium (2018)
13. Mohammad, S., Turney, P.: Crowdsourcing a word-emotion association lexicon. Computational Intelligence **29**, 436–465 (2013)
14. Odlin, T.: Language Transfer: cross-linguistic influence in language learning. Cambridge University Press, Cambridge, UK (1989)
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
16. Sanchez-Perez, M.A., Markov, I., Gómez-Adorno, H., Sidorov, G.: Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same Spanish news corpus. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. vol. 10456, pp. 145–151. Springer, Dublin, Ireland (2017)
17. Schmid, H.: Improvements In Part-of-Speech Tagging With an Application to German, pp. 13–25. Springer (1999)
18. Tsur, O., Rappoport, A.: 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. pp. 9–16. ACL, Stroudsburg, PA, USA (2007)