Portuguese Native Language Identification

Shervin Malmasi[1(✉)], Iria del Río[2], and Marcos Zampieri[3]

[1] Harvard Medical School, Boston, USA
shervin.malmasi@mq.edu.au
[2] University of Lisbon, Lisbon, Portugal
[3] University of Wolverhampton, Wolverhampton, UK

Abstract. This study presents the first Native Language Identification (NLI) study for L2 Portuguese. We used a sub-set of the NLI-PT dataset, containing texts written by speakers of five different native languages: Chinese, English, German, Italian, and Spanish. We explore the linguistic annotations available in NLI-PT to extract a range of (morpho-)syntactic features and apply NLI classification methods to predict the native language of the authors. The best results were obtained using an ensemble combination of the features, achieving 54.1% accuracy.

Keywords: Native Language Identification · Learner corpus Portuguese

1 Introduction

Native Language Identification (NLI) is the task of determining the native language (L1) of an author based on their second language (L2) linguistic productions [1]. NLI works by identifying language use patterns that are common to groups of speakers of the same native language. This process is underpinned by the presupposition that an author's L1, disposes them towards certain language production patterns in their L2, as influenced by their mother tongue. A major motivation for NLI is studying second language acquisition. NLI models can enable analysis of inter-L1 linguistic differences, allowing us to study the language learning process and develop L1-specific pedagogical methods and materials.

NLI research is conducted using learner corpora: collections of learner writing in an acquired language, annotated with metadata such as the author's L1 or proficiency. These datasets are the foundation of NLI experiments and their quality and availability has been a key issue since the earliest work in this area.

A notable research trend in recent years, and the focus of this paper, has been the extension of NLI to languages other than English [2]. Recent NLI studies on languages other than English include Chinese [3], Norwegian [4], and Arabic [5].

A. Villavicencio et al. (Eds.): PROPOR 2018, LNAI 11122, pp. 115–124, 2018.
https://doi.org/10.1007/978-3-319-99722-3_12

Since the learner corpus is a core component of NLI work, extending the task to a new language depends on the availability, or collection, of suitable learner corpora.

Early research focused on L2 English, as it is one of the most widely studied languages and data has been more readily available. However, continuing globalization has resulted in increased acquisition of languages other than English [6]. Additionally, researchers have sought to investigate whether the NLI methods that work for English would work for other languages, and whether similar performance trends hold across corpora. These motivations have led to an extension of NLI research to new non-English languages, of which our research directly contributes.

To the best of our knowledge, this study presents the first detailed NLI experiments on L2 Portuguese. A number of studies have been published on educational NLP applications and learner language resources for Portuguese, but so far none of them have included NLI. Examples of educational NLP studies that included Portuguese range from grammatical error correction [7] and automated essay scoring [8], to language resources such as the Portuguese Academic Wordlist (P-AWL) [9], and the learner corpus COPLE2 [10] which is part of the dataset used in our experiments.

The remainder of the paper is organized as follows: Sect. 2 discusses related work in NLI, Sect. 3 describes the methodology and dataset used in our experiments, and Sect. 4 presents the experimental results. Finally, Sect. 5 presents a brief discussion and concludes this paper with avenues for future research.

2 Related Work

NLI is a fairly recent, but rapidly growing area of research. While some research was conducted in the early 2000s, the most significant work has only appeared in recent years [11–15].

NLI is typically modeled as a supervised multi-class classification task. In this experimental design the individual writings of learners are used as training and testing data while the author's L1 information serves as class labels. NLI has received much attention in the research community over the past decade, with efforts focusing on improving classification [14], studying language transfer effects [16], and applying the linguistic features to other NLP tasks [17]. It has also been empirically demonstrated that NLI is a challenging task even for human experts, with machine learning approaches significantly outperforming humans on the same test data [18].

The very first shared task focusing on NLI was held in 2013, bringing further focus, interest and attention to the field.[1] The competition attracted entries from 29 teams. The winning entry for the shared task was that of [19], with an accuracy of 83.6%. The features used in this system are n-grams of words, parts-of-speech, as well as lemmas. In addition to normalizing each text to unit

[1] https://sites.google.com/site/nlisharedtask2013/home.

length, the authors applied a log-entropy weighting schema to the normalized values, which clearly improved the accuracy of the model. An L2-regularized SVM classifier was used to create a single-model system.

Growing interest led to another edition of the shared task in 2017, where the task was expanded to include speech data.[2] The results of the task showed that various types of multiple classifier systems, such as ensembles and meta-classifiers, achieved the best performance across the different tracks. While a number of participants attempted to utilize newer deep learning-based models and features (e.g. word embeddings), these approaches did not outperform traditional classification systems. Finally, it was also shown that as participants had used more sophisticated systems, results were on average substantially higher than in the previous edition of the task. A detailed report on the findings of the task can be found in [20].

With respect to classification features, NLI research has grown to use a wide range of syntactic, and more recently, lexical features to distinguish the L1. A more detailed review of NLI methods is omitted here for brevity, but a comprehensive exposition of the methods can be found in [21,22]. Some of the most successful syntactic and lexical features used in previous work includes Adaptor Grammars (AG) [23], character n-grams [24], Function word unigrams and bigrams [25], Word and Lemma n-grams, CFG Production Rules [12], Penn Treebank (PTB) part-of-speech n-grams, RASP part-of-speech n-grams [25], Stanford Dependencies with POS transformations [14], and Tree Substitution Grammar (TSG) fragments [13].

NLI is now also moving towards using models based on these features to generate Second Language Acquisition (SLA) hypotheses. In [26] the authors approach this by using both L1 and L2 data to identify features exhibiting non-uniform usage in both datasets, using them to create lists of candidate transfer features. The authors of [16] propose a different methodology, using linear SVM weights to extract lists of overused and underused linguistic features per L1 group.

Most English NLI work has been done using two corpora. The *International Corpus of Learner English* [27] was widely used until recently, despite its shortcomings[3] being widely noted [28]. More recently, TOEFL11, the first corpus designed for NLI was released [29]. While it is the largest NLI dataset available, it only contains argumentative essays, limiting analyses to this genre.

An important trend has been the extension of NLI research to languages other than English [5,30]. Recently, [3] introduced the Jinan Chinese Learner Corpus [31] for NLI and their results indicate that feature performance may be similar across corpora and even L1-L2 pairs. Similarly, [4] also proposed using the ASK corpus [32] to conduct NLI research using L2 Norwegian data.

In this study we also follow this direction, presenting new experiments on L2 Portuguese. Other aspects of our work, such as the classification methodology and features, are largely based on the approaches discussed above.

[2] https://sites.google.com/site/nlisharedtask/home.

[3] The issues exist as the corpus was not designed specifically for NLI.

3 Data and Method

3.1 Data

We used a sub-set of the NLI-PT dataset [33] containing texts for five L1 groups: Chinese, English, German, Italian, and Spanish. We chose these five languages because they are the ones with the greatest number of texts in NLI-PT. The sub-set is balanced in terms of proficiency level by L1. The composition of our data is shown in Table 1.

Table 1. Distribution of the five L1s in the NLI-PT datasets in terms of texts, tokens, types, and type/token ratio (TTR).

L1	Texts	Tokens	Types	TTR
Chinese	215	50,750	6,238	0.12
English	215	49,169	6,480	0.13
German	215	52,131	6,690	0.13
Italian	215	51,171	6,814	0.13
Spanish	215	47,935	6,375	0.13
Total	1,075	251,156	32,597	0.13

Texts in NLI-PT are automatically annotated using available NLP tools at two levels: Part of Speech (POS) and syntax. There are two types of POS: a simple POS with only the type of word, and a fine-grained POS with type of word plus morphological features. Concerning syntactic information, texts are annotated with constituency and dependency representations. These annotations can be used as classification features.

3.2 Classification Models and Evaluation

In our experiments we utilize a standard multi-class classification approach. A linear Support Vector Machine [34] is used for classification and feature vectors are created using relative frequency values, in line with previous NLI research [21]. A single model is trained on each feature type to evaluate feature performance. We then combine all our features using a mean probability ensemble.[4]

Similar to the majority of previous NLI studies, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. In recent years this has become the accepted standard for reporting NLI results. For generating our folds we use randomized stratified cross-validation which aims to ensure that the proportion of classes within each partition is equal [35]. While accuracy is a suitable metric as the data classes are balanced in our corpus, we also report per-class precision, recall, and F1-scores. We also compare these results against a random baseline.

[4] More details about this approach can be found in [21].

3.3 Features

Previous work on NLI using datasets which are not controlled for L1 and topic [3,5] avoids using lexical features. Using only non-lexicalized features allows researchers to model syntactic differences between classes and avoid any topical cues. For the same reasons, we do not use lexical features (*e.g.* word n-grams) as NLI-PT is not topic balanced. While a detailed exposition of this issue is beyond the scope of this paper, a comprehensive discussion can be found in [1, p. 23].

We extract the following topic-independent feature types: function words, context-free grammar production rules, and POS tags, as outlined below.

Function words are topic-independent grammatical words such as prepositions, which indicate the relations between content words. They are known to be useful for NLI. Frequencies of 220 Portuguese function words[5] are extracted as features. We also make this list available as a resource.[6]

Context-free grammar production rules are the rules used to generate constituent parts of sentences, such as noun phrases.[7] These rules can be obtained by first generating constituent parses for sentences. The production rules, excluding lexicalizations, are then extracted and each rule is used as a single classification feature. These context-free phrase structure rules capture the overall structure of grammatical constructions and global syntactic patterns. They can also encode highly idiosyncratic constructions that are particular to an L1 group. They have previously been found to be useful for NLI [12]. Our dataset already includes parsed versions of the texts which we used to extract these features.

Part-of-Speech (POS) tags are linguistic categories (or word classes) assigned to words that signify their syntactic role. Basic categories include verbs, nouns and adjectives, but these can be expanded to include additional morpho-syntactic information. The assignment of such categories to words in a text adds a level of linguistic abstraction. Our dataset already includes POS tags and n-grams of size 1–3 are extracted as features. They capture preferences for word classes and their localized ordering patterns. Previous work, and our own experiments, demonstrates that sequences of order 4 or greater achieve lower accuracy, possibly due to data sparsity, so we did not include them.

4 Results

In this section we first report results by individual feature types in terms of accuracy. Subsequently we report the results obtained using all features in an ensemble combination. Finally, we look at the performance obtained by the best system for each L1 class.

[5] Like previous work, this also includes stop words.
[6] http://web.science.mq.edu.au/~smalmasi/data/pt-fw.txt.
[7] They are also known as Phrase Structure Rules or Production Rules.

We first report the results obtained using systems trained on different feature types. Results are presented in terms of accuracy in Table 2. These results are compared against a uniform random baseline of 20%.

Table 2. Classification results under 10 fold cross-validation (accuracy is reported).

Feature type	Accuracy (%)
Random baseline	20.0
Function words	38.5
POS 1-grams	46.3
POS 2-grams	52.8
POS 3-grams	44.9
CFG production rules	43.3
Ensemble combination	54.1

We observed that all features types individually deliver results well above the baseline. POS bigrams are the features that individually obtain the best performance, achieving 52.8% accuracy. This demonstrates the importance of syntactic differences between the L1 groups. The ensemble combination, using all feature types, obtains performance higher than POS bigrams achieving 54.1% accuracy. These trends are very similar to previous research using similar features, but the boost provided by the ensemble is more modest. This is likely because the syntactic features used here are not as diverse as including other syntactic and lexical features, as shown in [36].

We also experimented with tuning the regularization hyperparameter of the SVM mode. This parameter (C) is considered to be the inverse of regularization strength; increasing it decreases regularization and vice versa. The results from the POS bigram model are shown in Fig. 1. We performed a grid search of the parameter space in the range of 10^{-6} to 10^{1}. We observe that model generalization (i.e. cross-validation score) is quite poor with strong regularization and improves as the parameter is relaxed. Generalization plateaus as approximately $C = 1$ and we therefore select this parameter value. Similar patterns hold for all feature types, but results are not included for reasons of space.

In Table 3 we present the results obtained for each L1 in terms of precision, recall, and F1 score as well as the average results on the five classes. Across all classes, we obtain a micro-averaged F1 score of 0.531 and a macro-averaged F1 score of 0.530.

Looking at individual classes, the results obtained for Chinese are higher than those of other L1s. One hypothesis is that as English, German, Italian, and Spanish are Indo-European languages, properties of Chinese, which belongs to the Sino-Tibetan family, are helping the system to discriminate Chinese texts with much higher performance than the other three L1s. To visualize these results

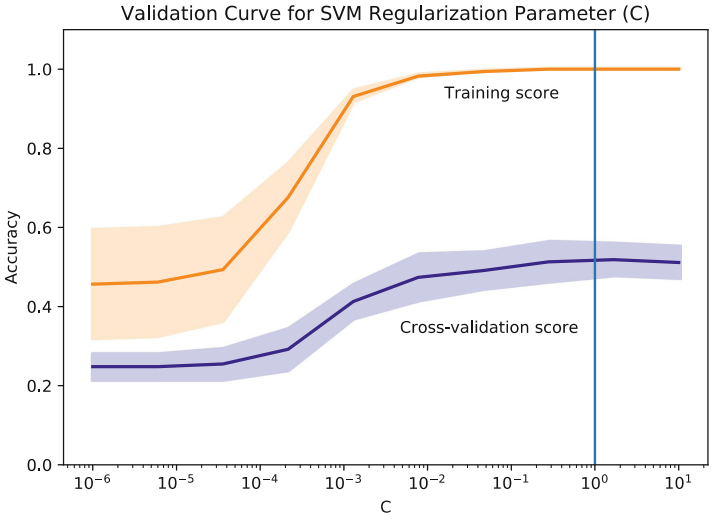

Fig. 1. Results for tuning the regularization hyperparameter (C) of the POS bigram SVM model. The top represents performance on the training set, while the bottom line is the cross-validation accuracy. The vertical line represents the value of C = 1.

and any notable error patterns, in Fig. 2 we present a heatmap confusion matrix of the classification errors.

Table 3. Ensemble system per-class results: precision, recall and the F1-score are reported.

Class	Precision	Recall	F1-score
CHI	0.571	0.796	0.665
ENG	0.507	0.326	0.397
GER	0.542	0.547	0.545
ITA	0.549	0.577	0.562
SPA	0.510	0.460	0.484
Average	0.536	0.541	0.531

Finally, another important finding here is that our results suggest the existence of syntactic differences between the L1 groups. Earlier in Sect. 3.3 we justified the use of non-lexical features to avoid topic bias, and the presence of such bias is also evidenced by the difference between our results and the lexical baseline provided with the dataset description [33]. Such lexical models built using topic-imbalanced datasets may not capture actual L1 differences between the classes. Accordingly, the results are often artificially inflated and may actually represent thematic classification.

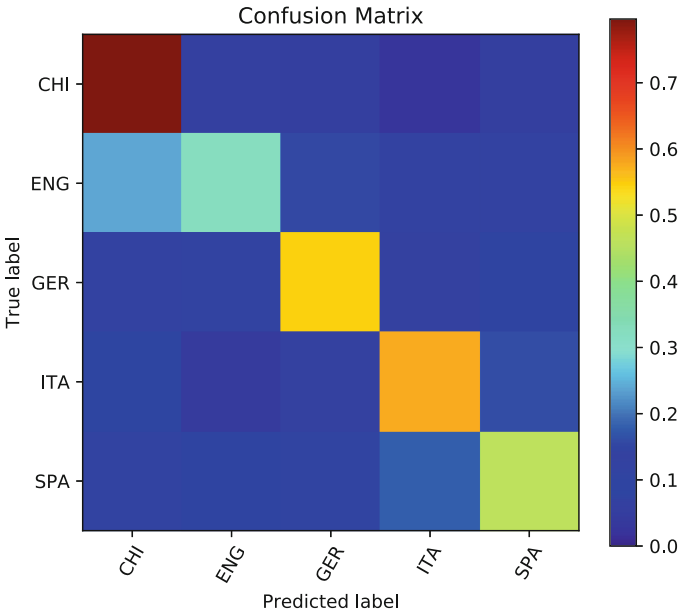

Fig. 2. Confusion matrix for our ensemble system.

5 Conclusion and Future Work

This paper presented the first NLI experiments on Portuguese. These results add to the growing body of evidence that demonstrates the applicability of NLI methods to various languages. The availability of the presented dataset also allows future research and hypotheses to be tested on another NLI corpus, which are valuable resources.

The presented results are comparable to those of other NLI studies [2], but not as high as those on the largest and most balanced corpora [20]. This is likely a limitation of our data, which we will address below.

This study opens several avenues for future research. One of them is investigating the influence of L1 in Portuguese second language acquisition. Such approaches, similar to those applied to English learner data [16], can have direct pedagogical implications. For example, the identification of the most discriminative language transfer features can lead to recommendations for language teaching and assessment methods. Such NLI models can provide the means to perform qualitative studies of the distinctive characteristics of each L1 group, allowing these differences to be described. Following this, further analysis may attempt to trace the linguistic phenomena to causal features of the L1 in order to explain their manifestation.

There are several directions for future work. The evaluation of more features, such as dependency parses, could be helpful. The application of more advanced ensemble methods, such as meta-classification [21], have also proven to be useful

for NLI, as well as other tasks [37,38]. However, we believe that the most valuable (and challenging) next step is the refinement and extension of the learner corpus. Having more data is extremely important in improving NIL accuracy. Additionally, well-balanced data is a key component of NLI experiments and having a dataset that is more carefully balanced for topic and proficiency will be of utmost importance for future research in this area.

Acknowledgements. We would like to thank the anonymous reviewers for the suggestions and constructive feedback provided.

References

1. Malmasi, S.: Native language identification: explorations and applications. Ph.D. thesis (2016)
2. Malmasi, S., Dras, M.: Multilingual native language identification. In: Natural Language Engineering (2015)
3. Malmasi, S., Dras, M.: Chinese native language identification. In: Proceedings of EACL. Association for Computational Linguistics, Gothenburg (2014)
4. Malmasi, S., Dras, M., Temnikova, I.: Norwegian native language identification. In: Proceedings of RANLP, Hissar, Bulgaria, pp. 404–412, September 2015
5. Malmasi, S., Dras, M.: Arabic native language identification. In: Proceedings of the Arabic Natural Language Processing Workshop (2014)
6. Block, D., Cameron, D.: Globalization and Language Teaching. Routledge, Abingdon (2002)
7. Martins, R.T., Hasegawa, R., Nunes, M.G.V., Montilha, G., De Oliveira, O.N.: Linguistic issues in the development of ReGra: a grammar checker for Brazilian Portuguese. Nat. Lang. Eng. **4**(4), 287–307 (1998)
8. Elliot, S.: IntelliMetric: From here to validity. In: A Cross-Disciplinary Perspective, Automated Essay Scoring, pp. 71–86 (2003)
9. Baptista, J., Costa, N., Guerra, J., Zampieri, M., Cabral, M., Mamede, N.: P-AWL: academic word list for Portuguese. In: Pardo, T.A.S., Branco, A., Klautau, A., Vieira, R., de Lima, V.L.S. (eds.) PROPOR 2010. LNCS (LNAI), vol. 6001, pp. 120–123. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12320-7_15
10. Mendes, A., Antunes, S., Janssen, M., Gonçalves, A.: The COPLE2 corpus: a learner corpus for Portuguese. In: Proceedings of LREC (2016)
11. Wong, S.M.J., Dras, M.: Contrastive analysis and native language identification. In: Proceedings of ALTA, Sydney, Australia, pp. 53–61, December 2009
12. Wong, S.M.J., Dras, M.: Exploiting parse structures for native language identification. In: Proceedings of EMNLP (2011)
13. Swanson, B., Charniak, E.: Native language detection with tree substitution grammars. In: Proceedings of ACL, Jeju Island, Korea, pp. 193–197, July 2012
14. Tetreault, J., Blanchard, D., Cahill, A., Chodorow, M.: Native tongues, lost and found: resources and empirical evaluations in native language identification. In: Proceedings of COLING, Mumbai, India, pp. 2585–2602 (2012)
15. Gebre, B.G., Zampieri, M., Wittenburg, P., Heskes, T.: Improving native language identification with TF-IDF weighting. In: Proceedings of BEA (2013)
16. Malmasi, S., Dras, M.: Language transfer hypotheses with linear SVM weights. In: Proceedings of EMNLP, pp. 1385–1390 (2014)

17. Malmasi, S., Dras, M., Johnson, M., Du, L., Wolska, M.: Unsupervised text segmentation based on native language characteristics. In: Proceedings of ACL (2017)
18. Malmasi, S., Tetreault, J., Dras, M.: Oracle and human baselines for native language identification. In: Proceedings of BEA (2015)
19. Jarvis, S., Bestgen, Y., Pepper, S.: Maximizing classification accuracy in native language identification. In: Proceedings of BEA (2013)
20. Malmasi, S., et al.: A report on the 2017 native language identification shared task. In: Proceedings of BEA (2017)
21. Malmasi, S., Dras, M.: Native Language Identification using Stacked Generalization. arXiv preprint arXiv:1703.06541 (2017)
22. Malmasi, S., Dras, M.: Native language identification with classifier stacking and ensembles. Computational Linguistics (2018)
23. Wong, S.M.J., Dras, M., Johnson, M.: Exploring adaptor grammars for native language identification. In: Proceedings of EMNLP (2012)
24. 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 (2007)
25. Malmasi, S., Wong, S.M.J., Dras, M.: NLI shared task 2013: MQ submission. In: Proceedings of BEA (2013)
26. Swanson, B., Charniak, E.: Data driven language transfer hypotheses. EACL **2014**, 169 (2014)
27. Granger, S., Dagneaux, E., Meunier, F., Paquot, M.: International Corpus of Learner English (Version 2). Presses Universitaires de Louvain, Louvian-la-Neuve (2009)
28. Brooke, J., Hirst, G.: Measuring interlanguage: native language identification with L1-influence metrics. In: Proceedings of LREC (2012)
29. Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., Chodorow, M.: TOEFL11: a corpus of non-native English. Educational Testing Service, Technical report (2013)
30. Malmasi, S., Dras, M.: Finnish native language identification. In: Proceedings of ALTA, Melbourne, Australia, pp. 139–144 (2014)
31. Wang, M., Malmasi, S., Huang, M.: The Jinan Chinese learner corpus. In: Proceedings of BEA (2015)
32. Tenfjord, K., Meurer, P., Hofland, K.: The ASK corpus: a language learner corpus of Norwegian as a second language. In: Proceedings of LREC (2006)
33. del Río, I., Zampieri, M., Malmasi, S.: A Portuguese native language identification dataset. In: Proceedings of BEA (2018)
34. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
35. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI **14**, 1137–1145 (1995)
36. Malmasi, S., Cahill, A.: Measuring feature diversity in native language identification. In: Proceedings of BEA (2015)
37. Malmasi, S., Dras, M., Zampieri, M.: LTG at SemEval-2016 Task 11: complex word identification with classifier ensembles. In: Proceedings of SemEval (2016)
38. Malmasi, S., Zampieri, M., Dras, M.: Predicting post severity in mental health forums. In: Proceedings of CLPsych (2016)