Supervised Identification of Writer's Native Language Based on Their English Word Usage

Agnieszka Jastrzebska
Faculty of Mathematics and Information Science
Warsaw University of Technology
Warsaw, Poland *A.Jastrzebska@mini.pw.edu.pl*

Wladyslaw Homenda
Faculty of Mathematics and Information Science
Warsaw University of Technology
and Faculty of Applied Information Technology
University of Information Technology and Management
Rzeszow, Poland *homenda@mini.pw.edu.pl*

Abstract

In this paper, we investigate the possibility of constructing an automated tool for the writer's first language detection based on a document written in their second language. Since English is the contemporary lingua franca, commonly used by non-native speakers, we have chosen it to be the second language to study. In this paper, we examine English texts from computer science, a field related to mathematics. More generally, we wanted to study texts from a domain that operates with formal rules. We were able to achieve a high classification rate, about 90%, using a relatively simple model (n-grams with logistic regression). We trained the model to distinguish twelve nationality groups/first languages based on our dataset. The classification mechanism was implemented using logistic regression with L1 regularisation, which performed well with sparse document-term data table. The experiment proved that we can use vocabulary alone to detect the first language with high accuracy.

Keywords: Classification, Natural Language Processing, First Language Identification, Second Language

1. Introduction

English has become the language of business and science. This has greatly benefited native speakers of English and encouraged everyone else to learn it. Although it is not easy to give exact numbers, it is estimated that there are about 350 million people who speak English as a first language and 1.5 billion people who speak English as a second language [6]. As we can see, there are more non-native than native speakers of English. It is also worth noting that English comes in different regional varieties, such as British English and Australian English. In this paper, we would like to explore the possibility of automatic detection of a person's first language based on based on their use of English words. The research focuses on written documents, and we use a number of supervised learning tools to recognize the first language.

While a trained ear may be able to recognize a speaker's first language based on their accent in English, this is much more difficult for written documents. When we listen to someone, we register how certain phonemes sound. The spoken words, in addition to their meaning, reveal a lot of information that may include the gender, age, and emotions of the speaker. In contrast, a written message conveys less "auxiliary" knowledge than a spoken message. On the other hand, written documents are much more convenient to obtain, store, and process than voice recordings. Therefore, using written documents as a medium for knowledge extraction is

a more practical strategy. The potential application areas of a first language identification cover a wide spectrum of fields, including personalized marketing and forensics.

Of course, this discussion can be generalized to languages other than English. The general problem addressed in this paper is the ability to automatically detect a first language based on word usage in a second language. This paper focuses on English, the most vibrant example of a second language used in the world.

From the point of view of information systems science, which is in the center of the ISD conference, this paper address selected issues in knowledge representation and reasoning area. We are working on new concepts and methods for knowledge extraction from text data.

Research Goal Formulation

The goal of the research undertaken in this paper was to provide a machine-learning-based method for first language recognition based on the vocabulary usage in English. We describe an empirical case study concerning a collection of 6,407 documents authored by 12 groups of subjects. Each group consisted of speakers we believe shared the same first language. We trained a supervised classification model that is able to recognize these groups based on a bag-of-words representation of the data.

Although native language identification has already been studied in the field of natural language processing, there are still some important aspects of this task that need to be investigated. An example of recent research in this direction is the work of Abdul-Mageed et al. in which a novel microdialect detection task was investigated based on a manually labeled dataset of 319 urban Arabic language microdialects [1]. Ionescu and Butnaru described a string kernels adaptation for polarity classification and Arabic dialect identification that achieves high accuracy rates [11]. Goldin et al. described a system that can accurately identify the native language of fluent non-native speakers based on a Reddit corpus [8]. This study is one of the first researches to accomplish the task of native language identification in a large-scale user-generated content scenario.

In contrast to existing research in this area, our work addresses a method for native language identification from documents extracted from a rather formal domain, namely computer science. In general, it is an example of a discipline that operates with formal rules. This goal accounts for the novelty/originality of this study. We use n-grams obtained from scientific texts. Let us recall that an n-gram is a sequence of n consecutive words extracted from a given corpus. If $n = 1$, then it is a single word, if $n = 2$, then it is a pair of words, etc. n-grams are simple but effective tools capable of capturing terminology. In our study, we considered $n = 1, 2, 3, 4, 5$. We were able to achieve high classification rates using a relatively simple model (n-gram + logistic classifier). We demonstrate a highly effective scheme for recognizing a writer's first language based on their word usage in a second language.

Limitations of the Study

We must emphasize that the problem of identifying the first language is very nuanced and complex. We wanted to bring new approach and ideas to solve this problem, but of course we do not aim to capture all the details. We had to impose some constraints, or rather assumptions, the main one being that we only used non-sensitive, publicly available information. This means that details such as third language proficiency, multilingual marriages and families, and many other aspects are not addressed in this study.

Contribution

An important contribution of this paper is a dataset that we created for this study and we are making available in a preprocessed form alongside this paper[1]. It consists of 6,407 documents (scientific texts) written in English by authors whose first language was American English, British English, Chinese, French, German, Italian, Japanese, Polish, Russian, Spanish, Turkish, and Vietnamese. We worked with a limited dataset. Our goal was to select a few relatively homogeneous counties in the sense of the languages effectively spoken. Perhaps the most challenging country in this dataset is China (due to the diversity of the languages spoken). The dataset we created, or more generally the methodology for material collection, was influenced by obvious budget constraints. Nevertheless, we do not know of any other publicly available dataset, so our contribution in this aspect is important and novel. It should be mentioned that the creation of a better dataset would require the processing of sensitive data of people of different nationalities (to establish nationality in a precise way), and the legal and practical organization of such an undertaking requires a tremendous effort. So our dataset, though imperfect, contributes value to this field.

Finally, we would like to mention that the proposed approach can support automatic authorship validation even though it does not solve this problem directly. Nonetheless, it can provide additional information about an author. Authorship validation is identified as an important practical task. Example study areas concern issues such as plagiarism detection, ghost-writing, deepfake or bot detection. Furthermore, other researchers point out that similar methods can find application in studies on second language acquisition and improvements of educational programmes [3].

The remainder of this paper is structured as follows. Section 2 discusses relevant literature positions. In Section 3, we present the methodology of the experiment. Section 4 discusses empirical experiments and their results. Section 5 concludes the paper.

2. Literature Review

In the context of the research presented in this article, it is worth mentioning that document analysis can be carried out at different levels. The most elementary level of processing is the character level. In some works, including the work of Ionescu et al. the authors postulate that we can use single characters and substrings to detect the first language of the writer [12]. In the cited paper, the authors define features (predictors) that either reflect different types of spelling errors or only certain types of words, such as function words. These features prove to be very effective in specific tasks. In the work cited above, string kernels were used to embed documents into a very large feature space delimited by all substrings of a certain length. The method used learning schemes and performed the task of feature selection and classification. The most important property of this approach is that the system operates at the character level, making it completely language independent.

Malmasi et al. have shown that Convolutional Neural Networks can be used to classify non-native speakers of a language [16]. Convolutional Neural Networks were applied at both the character level and the substring (and word) level. The alphabet consisted of characters - ‘a’ to ‘z’, ‘A’ to ‘Z’, and digits 0-9. To capture special discriminative features, they used convolution filters of different sizes. They experimented with different number of filters per filter size and maxpooling layers. They observed that larger filter sizes work well for this model. Malmasi et al. explain that this is because filters of size 6 or 7 attempt to capture whole words, while even larger filters capture word interactions [16]. Note, however, that Cruz-Uribe reports that the average length of an English word is 4.5 letters [4]. Based on the cited studies, one

[1]The entire dataset of papers in pdf format is available upon email request.

can conclude that it makes more sense to develop a word-level processing scheme rather than a character-level one. Nonetheless, it should be noted that the average accuracy of this model on the Educational Testing Service dataset was only about 30%. Additionally, this method is computationally demanding due to the use of Convolutional Neural Network architecture. Let us also mention that earlier work by the same authors [15], [17] laid the foundation for research on native language identification, including a discussion of human perception of foreign language use.

Next, let us address the study by Zampieri et al. in which Support Vector Machine-based ensembles were used for native language detection [23]. The idea behind classification ensembles is to improve overall performance by combining the results of multiple classifiers. Such systems have proven to be effective not only in native language and dialect identification, but also in numerous text classification tasks, among which we can mention the identification of complex words and the diagnosis of grammatical errors [18]. In the system described by Zampieri et al., classifiers used a variety of features [23]. The authors experimented with the following predictors: character n-grams (with n in $\{1, \ldots, 10\}$), word n-grams (with n in $\{1, 2\}$) extracted from essays and speech transcripts, and iVectors (iVectors are used to represent utterances in spoken language analysis). For the n-gram features, TF-IDF weighting (TF-IDF stands for Term Frequency - Inverse Document Frequency) was applied. The training dataset consisted of 11,000 essays, orthographic transcriptions of 45-second English oral responses, and iVectors (1000 instances for each of the eleven native languages). Surprisingly, the plain n-gram character-level approach without the other features was only marginally worse. The authors report classification accuracy reaching 83.55%.

An interesting perspective on native language identification was presented some time ago by Kochmar, where the author presents a technique based on error detection [13]. This study assumes that speakers of the same foreign language to English make the same grammatical errors in writing. In contrast, Koppel et al. propose to consider stylistic features instead of syntactic features in order to detect the native language of a person writing in English [14]. They identified aspects such as repeated use of certain types of neologisms or unusual word usage that can be used to identify the native language.

Several machine learning competitions are also worth mentioning, including the one described by Schuller et al. [22]. In this competition, participants were asked to develop an algorithm that assesses the Degree of Nativeness that evaluates English language proficiency.

3. Experiment Methodology

3.1. The Data Set

A new dataset was created for the case study discussed in this paper. Since it was not possible to cover all native languages, it was decided to narrow the scope of the study. Documents written in English by authors whose first language was American English, British English, Chinese, French, German, Italian, Japanese, Polish, Russian, Spanish, Turkish, and Vietnamese were collected. In order to keep the corpus consistent, which is a common practice in natural language processing (as Calzada Perez points out - note), all documents were research papers assigned to the computer science category in the Web of Science collection. All papers were published in English. The first language of the authors was manually verified by looking at their websites and using common sense knowledge. We then used another group of individuals to validate the labels created.

If a paper was written by more than one author and they were of different nationalities, it was rejected. The minimum length of the paper was five pages and it had to be published in a journal (not in conference proceedings). Only up to three papers per author was admitted in order to not distort the corpus.

Table 1. The cardinality of items in each class in the corpus.

China	1017	USA	906	UK	742	Russia	557
Spain	552	Turkey	449	France	446	Poland	381
Japan	354	Germany	342	Vietnam	331	Italy	330

Table 1 presents the number of papers obtained for each group. The final collection consisted of 6,407 papers by researchers from 12 countries. The dataset is available upon email request providing that the applicant proves rights to download the papers from the Web of Science otherwise a list of references to all papers will be provided.

The data set was slightly imbalanced; the majority class consisted of 1017 documents, while the least numerous class contained 330 samples. The selection of native languages considered was deliberate. We have Vietnamese, which is a language with a low morpheme-to-word ratio. We have one group for Chinese, knowing that Standard Mandarin is the official national language of China and is widely spoken and accepted. Standard Mandarin has a moderately high ratio of morphemes per word, but almost no inflectional affixes. It is assumed that both Vietnamese and Mandarin are analytic languages. The other languages belong to the group of synthetic languages in which meaning is modified by attaching dependent morphemes to the root morpheme [19]. It is worth mentioning that we have several representatives of agglutinative languages in the dataset: Turkish and Japanese. Agglutinative languages are a subtype of synthetic languages with morphemes always clearly distinguishable from each other.

It is also worth highlighting that we have two types of English language: American and British. We added these languages to our dataset as *control groups*. We are fully aware that English is the first language for both of these groups, but we wanted to see if there were any additional regional relationships. Most importantly, we wanted to see how often people who use English as a second language would be classified into these two groups. This particular classification error may, for instance, indicate that some papers were written very fluently.

The data collection methodology described above ensured the collection of papers written by a variety of authors, while it was still possible to pose hypotheses about authors nationality. Other data sources considered (for example, social media) do not offer the same benefits. The choice of processing scientific papers allowed for the use of already available resources and allowed for certain validation of labeling. We are aware that this dataset is not perfect, but there is no other such dataset available in the Internet.

3.2. Document Parsing

The first step was to convert the pdf files into plain text documents. After parsing the pdfs, the following items were removed:

- headers and footers;
- list of references;
- acknowledgements, division of work sections, etc.;
- paper title, authors names and affiliations;
- equations.

The listed content may have had a negative effect on the correctness of the experiment because it may have contained named entities. Leaving these items could have incorrectly increased the recognition accuracy. Therefore, they were removed.

The most difficult part was to remove acknowledgments to funding institutions, which in many cases were mentioned not only in standard places but also in the main text. We removed all words containing geographic locations and references to nations.

We also removed words that were incorrectly processed by the pdf-handling package using heuristic rules, such as words with more than two repetitions of the same letter. In addition, we have removed words shorter than 4 characters and words not existing in the English dictionary.

We have removed numbers, punctuation marks, and special characters. All letters have been converted to lowercase. We did not use a standard stop word list because we believe that some of these words play a significant role in sentence construction, and thus can be used as a discriminative feature when trying to detect a person's first language.

3.3. Document-Term Matrix Construction

Our goal is to construct a document-term matrix. This is a frequently used corpus representation technique, in which each row corresponds to a document and each column to a term. The content of the document-term matrix reflects the frequency of occurrence of a term in a given document. In its simplest form, the document-term matrix is binary. Zero means that the term is not present in the given document. One means that the term is present. Alternatively, we can have an integer document-term matrix, where the values represent the number of occurrences of a given term in a given document. Still another form is to weight the importance of a given term. This representation, known as the $TFIDF$ (Term Frequency - Inverse Document Frequency), is computed as follows:

$$TFIDF = TF \cdot IDF \tag{1}$$

$$TF(w, \mathbf{d}) = \frac{\text{\# of occurrences of term } w \text{ in document } \mathbf{d}}{\text{\# of occur. of the most frequent word in } \mathbf{d}} \tag{2}$$

$$IDF(w, \mathcal{D}) = log \frac{|\mathcal{D}|}{1 + |\{\mathbf{d} \in \mathcal{D} : w \in \mathbf{d}\}|} \tag{3}$$

In the equations above, the following notation is used: w is a term, $\mathbf{d}$ is a document, a collection of terms, $\mathcal{D}$ is a corpus, a collection of documents, $|\mathcal{D}|$ is the total number of documents in corpus $\mathcal{D}$.

In Eq. (3), in the denominator, we see 1, which is added to avoid numerical problems with the division. $TFIDF$ is a frequently used weighting scheme in text mining tasks [21]. It must be mentioned that "term" in a document-term matrix usually means an n-gram.

The major problem with the document-term matrix representation is that constructed matrices are sparse [2]. Thus, further processing must take into account dedicated dimensionality reduction methods, such as Latent Semantic Indexing, [10]. Alternatively, one may consider processing algorithms (here classification algorithms) that are capable to select features on their own [5].

It is worth emphasising, that the preprocessing applied in our experiment can be easily transferred to any other data set (in which the studied the second language is not in English, but, say, German).

3.4. Proof-of-Concept Model

To test how our models would potentially work on preprocessed data, we have constructed a preliminary proof-of-concept model. It was a random forest consisting of 1000 decision trees. It was trained using 75% randomly selected articles. Bigrams (i.e., n-grams of $n = 2$) were used as explanatory variables. The remaining samples served as the test set. Random forest classifier is a state-of-the-art method that is often used to solve classification problems in many domains of science [20].

We used two indicators to assess the quality of the model's fit: accuracy and F1-score:

$$\text{accuracy} = \frac{TP + TN}{ALL} \tag{4}$$

$$\text{F1-score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

where precision $= TP/(TP + FP)$ and recall $= TP/(TP + FN)$

TP (True Positives) is the number of correctly classified items from a given class, FP is the number of samples outside of a given class that were incorrectly classified to a given class. FN is the number of samples from a given class incorrectly classified to other classes. ALL is the number of all samples part taking in the experiment.

The proof-of-concept model obtained F1-score of 0.68 and accuracy of 0.75. These results are satisfactory. Furthermore, we tested the differences between the results achieved using $TFIDF$ and plain counts. The differences in these two quality measures were not statistically significant.

3.5. Main Experiment Setup

We chose two classifiers, a neural network with one hidden layer and categorical logistic regression with L1 regularization, as target classification methods. However, we quickly found that the best results on the test set were obtained using the logistic classifier. In the main experiment, all models were trained on a randomly selected 50% of all documents. The selection was purely random; class representation was not considered in the randomization procedure.

Logistic regression is a regression model capable of including categorical explanatory variables, such as class labels in our dataset [7]. When combined with LASSO (Least Absolute Shrinkage and Selection Operator), it allows us to regularize and select variables (this is L1 regularization).

The models based on artificial neural networks were fairly simple. Networks had one hidden layer with 8 neurons. We used two datasets: a dataset of terms consisting of unigrams and a dataset consisting of unigrams and bigrams. The number of training iterations was relatively low: 100. To avoid overfitting (when the model adapts too well to the training data), we used 5-fold cross-validation.

The categorical logistic regression model with L1 regularization was much better. So we spent more time experimenting with it. We tested several datasets:

- dataset made exclusively of unigrams;
- dataset consisting of unigrams and bigrams;
- dataset made of unigrams, bigrams, and trigrams;
- dataset with unigrams up to quintagrams.

We also experimented with different document-term matrix representations: count-based (default) and $TFIDF$. Besides, we tested whether it is beneficial to apply word stemming (we did not use it by default).

All experiments were performed on a mid-range Windows-based PC. The code was prepared in Python using sklearn for essential algorithms and pdfminer for pdf-operations. The computational cost of the trained models is low due to the nature of the algorithms we used. We see this as an advantage of the described processing scheme.

4. Results

4.1. Comparison of Results

In this section, we compare the following models:

- the proof-of-concept model with random forest and unigrams (marked rf-u),
- neural network with unigrams (marked as nn-u),

- logistic regression with:
  - unigrams (marked u),
  - unigrams and bigrams (marked ub),
  - unigrams, bigrams, and trigrams (marked ubt),
  - stemmed: unigrams, bigrams, and trigrams (marked ubt-s),
  - unigrams, bigrams, and trigrams, document-term matrix was weighted using the $TFIDF$ technique (marked ubt-t),
  - stemmed: unigrams, bigrams, and trigrams; document-term matrix was weighted using the $TFIDF$ technique (marked ubt-st),
  - unigrams up to quintagrams (marked ubtqq).

In Figure 1, we show the F1-score and the accuracy achieved by the tested models. The entire data set, summarized in Table 1, was randomly divided in half into training and test sets. The test set was not involved in training the model. The plot concerns the test set.

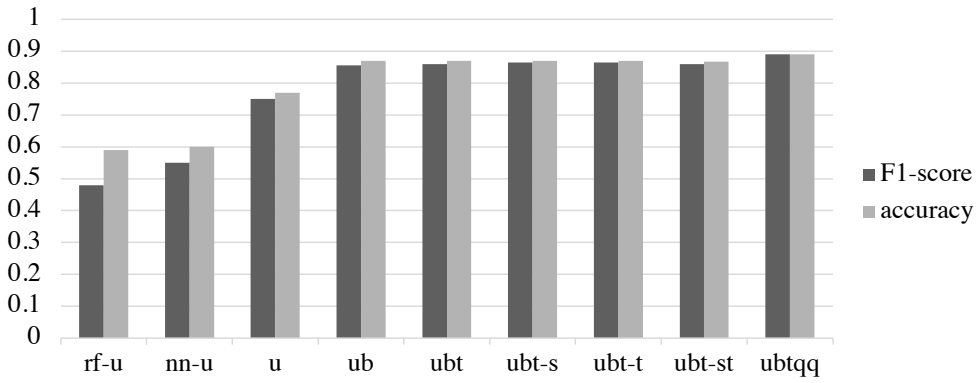

Fig. 1. Comparison of F1-score and accuracy for different experimental configurations. On the horizontal axis, we place short symbols denoting a particular configuration.

The best results were obtained by logistic regression with L1 regularization trained on n-grams of $n = 1, 2, 3, 4, 5$. The inclusion of longer n-grams allowed to construct better models. This can be seen in the increase in quality measures with an increase from model "u" to "ub", then a slight increase from "ub" to "ubt" and a slight increase from "ubt" to "ubtqq". Unsurprisingly, the largest improvement was obtained when bigrams were added to unigrams.

The models based on neural network and random forest were significantly worse than the model based on logistic regression. The sparsity of document-term matrices hinders the ability of a straightforward classification. Regularization therefore pruned weak predictors and resulted in much better models.

F1-score and accuracy reach similar values, indicating that the training procedures themselves were satisfactory.

Neither stemming nor the $TFIDF$ matrix representation were required to provide a good fit. They increased the computational complexity of the processing stream, but did not pay off with significant improvements in quality measures.

4.2. The Best Choice

We believe that the optimal choice is the "ubt" model, which is a logistic regression with L1 regularization constructed on unigrams, bigrams, and trigrams. It outputs reasonable classification quality while requiring relatively modest computational effort. Extensions to this model,

such as stemming or computing the $TFIDF$ representation, did not significantly improve the results.

In Table 2, we present the confusion matrix for the test set obtained with the "ubt" model. The F1-score for this model is equal to 0.86, while the accuracy is equal to 0.87.

Table 2. Confusion matrix for the test set. Full country names were replaced with abbreviations (for a pleasant layout). CN - China, FR - France, DE - Germany, IT - Italy, JP - Japan, PL - Poland, RU - Russia, ES - Spain, TR - Turkey, VN - Vietnam. Rows concern true classes, columns predicted classes.

	CN	DE	ES	FR	IT	JP	PL	RU	TR	UK	US	VN
CN	473	0	0	0	2	4	0	5	6	2	5	6
DE	4	131	4	2	2	1	4	5	2	2	12	0
ES	0	0	258	6	5	1	3	0	0	0	2	0
FR	2	3	4	195	4	0	2	4	0	4	5	0
IT	0	2	1	1	148	1	0	5	0	0	3	0
JP	10	4	0	1	0	143	3	9	3	1	7	2
PL	1	4	0	2	1	2	160	24	2	0	4	4
RU	0	1	1	2	0	1	3	277	0	0	0	1
TR	3	0	0	3	2	1	0	3	190	0	5	1
UK	7	9	4	2	3	1	4	4	9	311	17	1
US	8	9	2	5	6	11	0	9	7	23	371	7
VN	9	4	0	0	4	4	1	7	3	0	4	126

It is worth noting, that the biggest errors were made when the model discriminated American English from British English. We intentionally added these two groups as control groups to test their similarity. 24 papers written by Poles were classified to the Russian class. This is not surprising; both languages belong to the Slavic language family.

4.3. Interpretation of Results

Let us recall, that the purpose of the experiment was to test whether it is possible to detect person's first language from their document written in a second language (here English). It was addressed in a case study concerning 12 groups including two control groups.

In general, there were no problems with automatic discrimination of documents written by people whose first language was other than English. The most errors were made for the two control groups: US and UK.

We would like to point out an interesting result. A document written by a person whose first language was not English was more likely to be classified in the US class than in the UK class. We speculate that this may be because the subjects had more frequent contact with the American English than with the British English, but of course we have no evidence to support this suspicion.

The classification error leading to the assignment of a UK or USA label to a paper whose author is not a native English speaker can also be used as an indicator of language proficiency. Assuming that the model has been trained correctly, if a given document written by a non-native English speaker is classified to UK or USA, it means that the vocabulary of the non-native author resembles the vocabulary of a native English speaker. If this were true, although again no deeper studies were conducted in this regard, our experiment shows that Germans (ranked first) and Japanese (ranked second) in our dataset know English vocabulary best. At the same time, it is difficult to interpret the cases in which documents written by native English speakers were misclassified to some non-English speaking country.

We would like to note that the results of this experiment show that geographical proximity may induce the use of similar vocabulary and increase the chance of misclassification. For

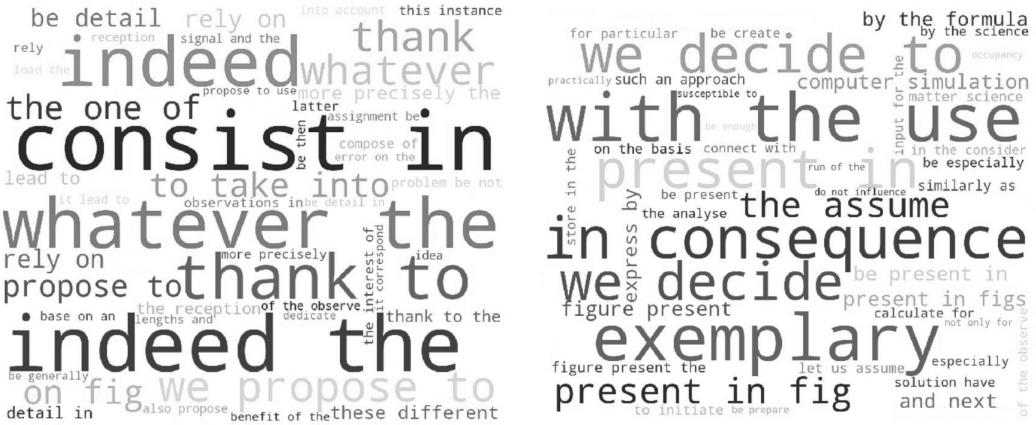

Fig. 2. Regression coefficients were used to plot word clouds of predictors (stemmed unigrams, bigrams and trigrams) that were the most important for class France (left) and class Poland (right). The size of a predictor corresponds to its weight.

example, Vietnamese were mostly misclassified into the Chinese or Russian class, even though the linguistic typology indicates that Vietnamese and Russian belong to different groups of languages. The same is true for the Japan class, which was most often misclassified into the US, China, or Russia class. This is a very far-fetched generalization, but we would like to mention an interesting study of Graessner et al. [9] showing that semantics is often region-dependent. A more plausible explanation for this is the use of similar expert terminology.

We would like to note, that the results of this experiment show that geographical closeness may induce the usage of similar vocabulary and increase the chance of misclassification. For instance, Vietnamese texts were most often incorrectly classified to the China class or Russian class, even though linguistic typology suggests that those two languages belong to different groups. The same concerns Japanese documents, that were most likely to be incorrectly classified to USA, China, or Russia class. This may be a very far-fetched generalisation, but we would like to mention that semantics is frequently region-dependent [9]. The most plausible explanation of this fact in the context of our study is that there may be a closer research collaboration in neighbouring regions. This, in turn, would result in publishing papers in very similar sub-fields of computer science. Thus, the usage of professional terminology might have been the reason why we observed this geographical dependence in our data set.

One of the many benefits of using a logistic regression algorithm is that we can use the fitted regression coefficients to compare the validity of different predictors. We used these coefficients to plot the word clouds.

The Figure 2 shows the word clouds of the predictors (unigrams, bigrams, and trigrams) that were most important in recognizing the France class (left image) and the Poland class (right image). The font size corresponds to the predictor's weight. We can use such word clouds to compare the impact of different predictors in the class assignment procedure. It was more intuitive for us to analyze word clouds based on stemmed vocabulary (graphs are clearer). Comparing the word clouds (stemmed words) for the classes France and Poland, we see, for example, that in the studied set of documents the French write "on fig" and the Poles write "in fig". Also, Poles use the word "present" very frequently. Further conclusions can be drawn if we plot the word clouds for the other groups.

There is always a risk that the obtained results are biased by the terminology used in the collected data samples. For example, in Figure 2, we see that French used "computer simulation" phrase often, which could be caused by a poor selection of papers which were too often on computer simulation. Still, we believe that the value of our study is rather on the conceptual

level, that we were able to use text data to infer knowledge about a seemingly distinct matter.

5. Conclusion

In this paper, we presented an efficient scheme for building a supervised model for recognizing a writer's first language from a document written in their second language. We designed an experiment in which we tested English as a second language. The results are very promising; we were able to achieve a classification accuracy of about 90%. The most errors were made in recognizing the group of papers originating from the UK and the US. These two groups, although not quite matching the purpose of this study, were added as a control set. Besides, we were able to draw interesting conclusions when we compared the classification accuracy for these two groups and the other groups.

It should be noted that the proposed model does not require the writer's knowledge of English to ensure class assignment. Moreover, the choice of knowledge representation makes the described approach transferable to research on other languages.

The proposed method uses vocabulary as predictors. In particular, n-gram are used, where $n = 1, 2, 3, 4, 5$. The n-grams convey elementary grammatical structures, but essentially, the proposed method is a bag-of-words vocabulary-based approach.

The most important advantage of the present work is that the model (n-gram + logistic classifier) is simple but effective for the task of native language identification in the field of computer science.

Future research directions arising from the presented study include answering the question of how much the task of native language identification depends on the domain of the documents being processed. Specifically, whether research papers in the mathematical sciences and, for example, the social sciences differ from each other from the point of view of native language identification. In other words, to what extent do domains of science with very formal rules affect second language use and mother tongue identification.

Acknowledgment

The project was funded by POB Research Centre for Artificial Intelligence and Robotics of Warsaw University of Technology within the Excellence Initiative Program - Research University (ID-UB).

References

1. Abdul-Mageed, M., Zhang, C., Elmadany, A., Ungary, L.: Toward micro-dialect identification in diaglossic and code-switched environments. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. pp. 5855–5876. Association for Computational Linguistics (2020)
2. Ailem, M., Role, F., Nadif, M.: Model-based co-clustering for the effective handling of sparse data. Pattern Recognition **72**, 108 – 122 (2017)
3. Bestgen, Y., Granger, S., Thewissen, J.: 5. Error Patterns and Automatic L1 Identification, pp. 127–153. Multilingual Matters (2012)
4. Cruz-Uribe, D.: Cryptography resources (2018), `https://www.cs.trincoll.edu/~crypto/resources/david.html`
5. Deng, X., Li, Y., Weng, J., Zhang, J.: Feature selection for text classification: A review. Multimedia Tools and Applications **78**(3), 3797–3816 (Feb 2019)
6. Eberhard, D.M., Simons, G.F., Fennig, C.D.: Ethnologue: Languages of the world. twenty-second edition (2019), `https://www.ethnologue.com`
7. Fernandez-Delgado, M., Sirsat, M., Cernadas, E., Alawadi, S., Barro, S., Febrero-Bande, M.: An extensive experimental survey of regression methods. Neural Networks

111, 11–34 (2019)
8. Goldin, G., Rabinovich, E., Wintner, S.: Native language identification with user generated content. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3591–3601. Association for Computational Linguistics, Brussels, Belgium (2018)
9. Graessner, A., Zaccarella, E., Hartwigsen, G.: Differential contributions of left-hemispheric language regions to basic semantic composition. Brain Structure and Function **226**(2), 501–518 (2021)
10. Horasan, F., Erbay, H., Varcin, F., Deniz, E.: Alternate low-rank matrix approximation in latent semantic analysis. Scientific Programming **2019** (2019)
11. Ionescu, R.T., Butnaru, A.: Improving the results of string kernels in sentiment analysis and arabic dialect identification by adapting them to your test set. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 1084–1090. Brussels, Belgium (2018)
12. Ionescu, R.T., Popescu, M., Cahill, A.: Can characters reveal your native language? a language-independent approach to native language identification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 1363–1373. Doha, Qatar (2014)
13. Kochmar, E.: Identification of a writer's native language by error analysis (June 2011), `https://www.cl.cam.ac.uk/ek358/Native_Language_Detection.pdf`
14. Koppel, M., Schler, J., Zigdon, K.: Automatically determining an anonymous author's native language. In: Kantor, P., Muresan, G., Roberts, F., Zeng, D.D., Wang, F.Y., Chen, H., Merkle, R.C. (eds.) Intelligence and Security Informatics. pp. 209–217. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
15. Malmasi, S.: Native language identification: Explorations and applications (2016)
16. Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., Napolitano, D., Qian, Y.: 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. pp. 62–75. Copenhagen, Denmark (2017)
17. Malmasi, S., Tetreault, J., Dras, M.: Oracle and human baselines for native language identification. In: Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications. pp. 172–178. Denver, Colorado (Jun 2015)
18. Moyano, J.M., Gibaja, E.L., Cios, K.J., Ventura, S.: Review of ensembles of multi-label classifiers: Models, experimental study and prospects. Information Fusion **44**, 33 – 45 (2018)
19. Nichols, J.: Linguistic Diversity in Space and Time. University of Chicago Press (1992)
20. Probst, P., Wright, M.N., Boulesteix, A.L.: Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**(3), e1301 (2019)
21. Sabbah, T., Selamat, A., Selamat, M.H., Al-Anzi, F.S., Viedma, E.H., Krejcar, O., Fujita, H.: Modified frequency-based term weighting schemes for text classification. Applied Soft Computing **58**, 193–206 (2017)
22. Schuller, B., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J., Baird, A., Elkins, A., Zhang, Y., Coutinho, E., Evanini, K.: The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH **08-12-September-2016**, 2001–2005 (2016)
23. Zampieri, M., Ciobanu, A.M., Dinu, L.P.: Native language identification on text and speech. CoRR **abs/1707.07182** (2017)