Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2020"

Moscow, June 17–20, 2020

NATIVE LANGUAGE IDENTIFICATION FOR RUSSIAN USING ERRORS TYPES

Remnev N. V. (nremnev@hse.ru)

National Research University Higher School of Economics,
Moscow, Russia

The task of recognizing the author's native (Native Language Identification—NLI) language based on a texts, written in a language that is non-native to the author—is the task of automatically recognizing native language (L1). The NLI task was studied in detail for the English language, and two shared tasks were conducted in 2013 and 2017, where TOEFL English essays and essay samples were used as data. There is also a small number of works where the NLI problem was solved for other languages. The NLI problem was investigated for Russian by Ladygina (2017) and Remnev (2019). This paper discusses the use of well-established approaches in the NLI Shared Task 2013 and 2017 competitions to solve the problem of recognizing the author's native language, as well as to recognize the type of speaker—learners of Russian or Heritage Russian speakers. Native language identification task is also solved based on the types of errors specific to different languages. This study is data-driven and is possible thanks to the Russian Learner Corpus developed by the Higher School of Economics (HSE) Learner Russian Research Group on the basis of which experiments are being conducted.

Keywords: native language identification, NLI, support vector machine, SVM, TF-IDF, Russian Learner Corpus

DOI: 10.28995/2075-7182-2020-19-1123-1133

ОПРЕДЕЛЕНИЕ РОДНОГО ЯЗЫКА АВТОРА ДЛЯ РУССКОГО ЯЗЫКА

Ремнев Н. В. (nremnev@hse.ru)
НИУ ВШЭ, Москва, Россия

1. Introduction

The task of native language identification (NLI) is a search for common linguistic patterns characteristic of a group of speakers of the same language, while the desired patterns should allow distinguishing groups of texts of speakers of one language from another. The possibility of finding these patterns is based on assumption that the author's native language has a definite impact on the language being studied. The described influence can be in a variety of forms: it can be a word order not specific for the language being studied; the use of words from the native language directly in written and oral speech; modifications of speech according to models of the native language and others.

The NLI Shared Task 2013 [Tetreault et al., 2013] and 2017 [Malmasi et al., 2017] competitions are based on the corpus of English essays from TOEFL exam, written by representatives of 11 countries. The results of the competition show that the classification according to the author's native language is quite possible: there is rather high percentage of accuracy (more than 80%), while patterns mentioned above are clearly visible. For example, for the Arabic language in TOEFL 11 case, incorrectly spelling of the word *alot* is often used (correct English version is *a lot*).

Texts written in a non-native language may belong to two groups of speakers. The first group is students of a foreign language. A little more complicated and interesting speaker type is the group of heritage speakers. As a rule, these are people who learned a language in childhood, but due to various reasons (most often it is emigration) they use another language as the main one. For these people, stronger intersection of languages is typical, as a result unique errors that are very rarely found in foreign language learners arise. As an example, we can name incorrect word formation: the word *добрость* is incorrectly derived from *добрый* (*kind*) by analogy with *злость* (*anger*) / *злой* (angry), while the correct Russian word for *kindness* is *доброта*.

This paper presents an approach to recognizing the author's native language by his or her texts written in Russian, and also addresses the issue of classifying texts by speaker type. For this purpose, we use the Russian Learner Corpus maintained by the HSE Learner Russian Research Group [Rakhilina et al., 2016]. This corpus, in addition to the texts and labels of the author's native language, contains a label of the speaker's type: learner of Russian or Heritage Russian speaker and manually marked errors made by authors with error type label. In this work, native language identification based on errors specific to different languages is also object of the research.

The task of recognizing the author's native language has several applications in various fields. Thus, the task may be useful for identifying the author [Estival

et al., 2007], which is necessary when conducting forensic examinations [Gibbons, 2003]. Of course, the NLI task is also useful for studying interference [Malmasi and Dras, 2014] and other kinds of linguistic studies, as it allows highlighting interesting patterns that are difficult to identify in a particular text. Another area of application is language teaching: knowing typical errors or unusual language constructs that are identified by solving the NLI task can be useful for adapting the language learning process for different language groups with a common mother tongue [Rozovskaya and Roth, 2011] through encouraging a student to focus on certain common mistakes when learning a language.

This paper is organized as follows: the next section briefly summarizes the results of previous NLI studies with the main focus on the results of the NLI Shared Task competitions of 2013 and 2017. After that, a sufficiently detailed description of Russian Learner Corpus is given. The next section presents experimental results obtained on the Russian Learner Corpus for the tasks of identifying the author's native language and speaker type basing both on texts and errors types. We conclude by pointing out several directions for further research.

2. Related work

Most studies in the field of NLI are based on English texts and use both lexical and synthetic features. Popular lexical features include symbolic and vocabulary n-grams; and, among synthetic features, POS (part-of-speech) tags and others can be distinguished. The researchers often use support vector machines as a classification method; however, in general, the emphasis is put on the development and combination of features, rather than on the classification method. A detailed review of NLI state of the art (as of 2016) is presented in the Ph.D. thesis of S. Malmasi [Malmasi, 2016]. Here, we provide a brief overview of the topic development, as well as of several significant works written by the team leaders of the NLI Shared Task 2013 and 2017 competitions, so as to highlight the general trends in the use of classifiers and features.

For the 2013 and 2017 competitions [Tetreault et al., 2013], [Malmasi et al., 2017], which largely formed the NLI area and gave a significant impetus to the development of the theme, TOEFL 11 corpus, introduced in 2013 by [Daniel Blanchard et al., 2013], was used (in the 2017 competition, the corpus was extended with additional data). The corpus contains data from native speakers of eleven languages, 1100 essays for each language. In addition, the corpus contains metadata, such as English proficiency (low/medium/high) and the subject of the text (there are 8 different essay topics per corpus in total). TOEFL 11 corpus was specifically developed for the 2013 NLI Shared Task; since then, it has become a commonly used dataset to compare NLI models.

In the 2013 competition, the various feature sets based on n-grams of different orders (symbolic, vocabulary, and POS) are used by 29 participating teams; the most commonly used classification method was SVM, although some other methods were also used. The approximate distribution of classification methods according to the results of the competition [Tetreault et al., 2013] based on the works presented by 24 teams is demonstrated in **Table 1**.

Table 1. Distribution of methods used by teams

Machine Learning Method	Teams used count
SVM	14
Ensemble	4
Maximum Entropy / logistic regression	3
Discriminant Function Analysis	1
String Kernels / Local Rank Distance	1
PPM	1
kNN	1

The highest accuracy was achieved by Jarvis team [Jarvis et al., 2013]: it equals to 83.6%. The authors used a large number of features including symbolic, vocabulary and POS n-grams on which they trained a classifier using the support vector machine. The second most accurate result was shown by Oslo team—83.4% [Lynum, 2013], they also use SVM classifier trained on symbolic n-grams. Among the teams that did not use SVM, 82.7 % accuracy was reached: Unibuc team [Popescu and Ionescu 2013] used String Kernels and Local Rank Distance and MITER team used a combination of machine learning methods [Henderson et al., 2013]. Other teams that received high accuracy results used mainly support vector machine classifier with different features based on n-grams: [Bykh et al., 2013] used vocabulary 1..2-grams, POS 1..5-grams and dependency analysis; [Goutte et al., 2013]—vocabulary 1..2-grams, POS 2..4-grams and dependency analysis; [Gebre et al., 2013]—character 1..6-grams, vocabulary 1..2-grams and POS 1..4-grams; [Mizumoto et al., 2013]—character 2..3-grams, vocabulary 1..2-grams, POS 2..3-grams and analysis dependencies; [Wu et al., 2013]—vocabulary 1..2-grams. Due to works provided above, we can quite clearly observe tendencies in the use of classification methods and features.

The 2017 competition was divided into three areas: an essay, oral speech, and a combination of essay and oral speech. In total, 19 teams took part in the competition, 17 of which have published their works based on which a report was made [Malmasi et al., 2017]. The highest accuracy in the competition was achieved by the ItaliaNLP Lab team [Cimino and Dell'Orletta, 2017] which equals to 88.18%. The authors used a rather interesting approach combining the results of two classifiers. The first classifier based on logistic regression works at the sentences level; results are provided to the second classifier which uses the support vector method and already works at the level of the whole text. The second result (88.08%) was shown by the CIC-FBK team [Markov et al., 2017] also used the support vector machine based on the standard set of features, such as symbolic, vocabulary and POS n-grams, functional words. In addition to these features, several new features including syntactic n-grams were also used. The features were weighted using log-entropy. The interesting approach is also presented in the work of the NRC team [Goutte and Leger 2017]: the authors used about 10 SVM classifiers following by a vote to get the final result.

Drawing conclusions from the review of works in the field of NLI, we can talk about the dominance of various features based on n-grams (most often symbolic and vocabulary) due to the simplicity of their formation and rather high uniqueness indicators for various languages in the classification. SVM and a combination of several methods (usually several SVM classifiers) is the most common used among the classification methods.

3. Data description

The Russian Learner Corpus, presented by HSE Learner Russian Research Group under the direction of E. V. Rakhilina, is used as data in the work [Rakhilina et al., 2016]. This corpus contains samples of oral and written speech of two speaker types in Russian: those who study Russian as a foreign language and those who are heritage speakers. The main part of the corpus texts is provided by teachers of Russian as a foreign language abroad, many of which also work with heritage speakers. The corpus includes both academic and non-academic texts.

The corpus contains 15 languages, including Chinese, Danish, English, Estonian, Finnish, French, German, Italian, Japanese, Kazakh, Korean, Norwegian, Serbian, Swedish and Thai. The distribution of texts by language is presented in **Table 2**.

Table 2. Distribution of texts by languages

Language	Label	Count
Chinese	Chi	24
Danish	Dut	18
English	Eng	3,145
Estonian	Est	2
Finnish	Fin	1,231
French	Fr	495
Deutsch	Ger	284
Italian	Ita	115
Japanese	Jap	1,571
Kazakh	Kaz	494
Korean	Kor	197
Norwegian	Nor	28
Serbian	Ser	19
Swedish	Swe	178
Thai	Tai	30

Some languages, for example, Estonian and Danish, are represented by a small number of texts, while English, Finnish and Japanese have a rather large number of texts (more than 1,000). Such an imbalance requires either the exclusion of languages with a small number of texts, or, on the contrary, the addition of new texts by generating texts based on current ones, or the addition of a corpus with real texts of the required languages. The total number of texts with known language labels is 7,831, while the total number of corpus texts is more than 8,000.

Another important label of the text is the type of speaker—a student of the Russian language, or heritage speaker. A classifier developed to recognize the author's native language of a text is also used to classify by type of speaker. Total texts with defined label FL (foreign) or HL (heritage)—7,953. **Table 3** shows the distribution of corpus texts by these labels.

Table 3. Distribution of texts by speaker type

Speaker Type	Label	Count
Foreign	FL	5,519
Heritage	HL	2,434

In addition to the author's first language labels and the speaker's type, the corpus also contains metadata about errors which were identified and corrected. In dataset original text contain also corrected version and for each error identified correction with tag is provided. Errors classification in Russian Learner Corpus has a rather complex hierarchy. Common types of errors, such as spelling, morphological, syntax and errors in constructions are divided into several subtypes. In some cases, error could be identified with several error types. Table 4 shows the distribution of error tags for texts written by English speakers (3 most common error tags). In Table 4 "lex" stands for misuse of words, "ortho" for spelling errors and "const"—errors in constructions.

Table 4. Distribution of error tags for English texts

Error Tag	Count
lex	5,547
ortho	3,997
constr	3,270

Table 5 shows 3 most common error tags for each of the author native languages (only languages presented by more than 100 errors found are used).

Table 5. Top error tags by native language

Language	Top error tags
English	Lex, ortho, constr
Finnish	Ortho, lex, gov
French	Syntax, ortho, subst
Deutsch	Ortho, lex, subst
Japanese	Transfer, not-clear, lex
Kazakh	Lex, ortho, gov
Korean	Lex, punc, not-clear
Chinese	Transfer, par, gov
Italian	Constr, ortho, arggender
Norwegian	Gov, transfer, lex

Table 5 shows that each language differs from another in its set of errors. This statement allows to conclude that number of error tags of the text can help to unique determine author native language.

4. Results

To solve the problem of recognizing the author's native language for the Russian language, we develop the classification model based on the support vector method and the TF-IDF metric. SVM-TF-IDF approach is frequently used in NLP (NLI in particular) and is more applicable to this work rather than modern neural networks for texts approaches (including deep learning and convolutional/recurrent neural networks). The reason is the small number of texts in corpus used in this work. The data provided by Russian Learner Corpus is unique and generate texts close to original training data seems to be difficult to accomplish. In this section, we provide the results of testing the model on the data of the Russian Learner Corpus.

The corpus contains 7,831 texts, for which the native language of the author is given. Table 1 shows that for a higher quality of the model's work, it is necessary to balance the data of the corpus. For this, texts relating to the language for which the corpus contains less than 178 texts (the number of texts for the Swedish language) are excluded from the data set. Thus, the data for training and testing the model includes the following languages: English, Finnish, French, German, Japanese, Kazakh, Korean and Swedish. For each language, 178 texts are selected using the following rule: average texts size in each group of texts be close to 182 words (mean texts size for Swedish). The total number of texts was 1,424. The texts are divided into training and test samples in the ratio of 70% (996 texts) to 30% (428 texts). In this case, the splitting occurs in such a way that in the training and test samples there is an equal ratio by the number of texts by languages. The texts used were also processed by removing "quot", "gt", "lt" tags. Other importance preprocessing action was removing texts theme anomalies for obtaining proper results: for example, in Japanese authors texts number of essays were dedicated to ecology theme, containing special lexicon.

Table 6 presents the results of experiments for different sets of n-grams used in the model. As can be seen from the results, the maximum accuracy is achieved for small orders of n-grams: the higher order of n-grams leads to the lower resulting accuracy of the model.

Table 6. The results of the experiments for different N-grams settings

n-grams	Precision
1, 2	0.8025
1, 3	0.7919
1, 4	0.7619
2, 3	0.7242
2, 4	0.7136
3, 4	0.6461

In Table 7, for the first experiment from Table 6 (as an experiment in which maximum accuracy is achieved) a more complete interpretation of the classification results is presented.

Table 7. The results for classification for experiment configuration with maximum accuracy

	Precision	Recall	F1-score
Eng	0.78	0.86	0.82
Fin	0.73	0.60	0.66
Fr	0.65	0.90	0.75
Ger	0.67	0.81	0.74
Jap	1.00	0.93	0.96
Kaz	1.00	0.75	0.85
Kor	0.80	0.92	0.85
Swe	0.79	0.70	0.74
Micro avg.	0.80	0.80	0.80

High accuracy achieved for Kazakh language could be explained by some native language words used in texts, which are close to Russian in writing. This anomaly was missed during preprocessing. Another factor that could explain high accuracy is similarity with Russian—less number of mistakes for that reason. Japanese language was also classified with maximum precision. The most notable features for Japanese include words “очень”, “если”, “друг”. At least word “друг” in top features list could be explained by some number of Japanese authors texts written on the friendship theme.

Another task considered in this work is the classification of texts according to the type of speaker: learner of Russian or Heritage Russian speaker. Total number of texts with a well-known label of the speaker type is equal to 7,953. It can be noted that the data used by the model also contains an imbalance problem: FL label (learners of Russian) is contained in 5,519 texts, and the HL label (heritage speakers) is set to 2,434 texts. Principles of preprocessing for speaker type classification texts were similar to NLI classification including FL language authors texts selection was made with respect to native languages proportion in HL speakers. After balancing the data, the following results presented in Table 8 are obtained (model configuration is the same as for native language identification task).

Table 8. The results for the classification by speaker type

	Precision	Recall	F1-score
FL	0.84	0.90	0.87
HL	0.90	0.85	0.87
Micro avg.	0.87	0.87	0.87

Top features for FL include mixing alphabets errors like “Друг” word, while HL speakers often make morphological and new words creation errors like “дружбость”, “добрость” and “котошки” (коты/кошки). It is worth noting, that preprocessing was made with native languages factor, however imbalance was still remained and may affect accuracy of classification.

Both native language identification and speaker type classification were also solved based on error types. The core idea is to use each error tag appeared in dataset as a feature and represent each text using these features defined. For classification based on error types the model described above is employed. The following results are obtained: for the task of recognizing the native language of the author of the text for the Russian language the accuracy is equal to 80%, and for the task of determining the speaker type the accuracy is equal to 76%.

5. Conclusions and Future Work

This paper presents the results of a study to recognize the author's native language based on a Russian text using Russian Learner Corpus. The developed model based on the support vector method using the TF-IDF metric for the vector representation of the texts makes possible to achieve about 80% accuracy in classification. Another task, which is also considered within the framework of the work, is the classification according to the speaker type: learners of Russian or Heritage Russian learners. The developed model addresses this problem with a high accuracy of 87% achieved. It is worth noting several important points regarding some features of this work.

It is worth noting several important points regarding some features of this work. First, it is necessary to perceive the results tacking into account the data used. The model is trained on small amount of texts, since the corpus itself is not sufficiently balanced and there are significant differences in the number of texts for different languages; for example, Russian texts written by the English are about 6 times more than texts written by the French, while both of these languages are represented quite a large number of texts for the corpus. Another important point is the relative simplicity of the model, so it is logical and expected that higher classification accuracy can be achieved. The model is based on well-proven approaches in NLP and NLI in particular: in addition to the support vector method and the TF-IDF metrics, vocabulary n-grams of different orders are used.

References

1. *Ladygina* (2017), Native language identification for Russian using referential features. Master of Arts Thesis.
2. *Remnev N.* (2019), Native Language Identification for Russian. IEEE International Conference on Data Mining Workshops, ICDMW.
3. *Joel Tetreault, Daniel Blanchard, Aoife Cahill* (2013), A Report on the First Native Language Identification Shared Task. Educational Testing Service.
4. *Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, Yao Qian* (2017), A Report on the Native Language Identification Shared Task. Educational Testing Service.
5. *Ekaterina Rakhilina, Anastasia Vyrenkova, Elmira Mustakimova, Alina Ladygina, Ivan Smirnov* (2016), Building a learner corpus for Russian. In Proceedings of the

joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC.
6. *Dominique Estival, Tanja Gaustad, Son Bao Pham, Will Radford, Ben Hutchinson* (2007), Author profiling for english emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING), pages 263–272.
7. *Gibbons Pauline* (2003), Mediating language learning: teacher interactions with ESL students in a content-based classroom, TESOL Quarterly no. 37/2, pages 247–273.
8. *Shervin Malmasi, Mark Dras* (2014), Language Transfer Hypotheses with Linear SVM Weights. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14).
9. *Alla Rozovskaya, Dan Roth* (2011), Algorithm Selection and Model Adaptation for ESL Correction Tasks. In Prodeedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 19–24.
10. *Shervin Malmasi, Mark Dras* (2014), Chinese Native Language Identification. EACL.
11. *Shervin Malmasi, Mark Dras* (2015), Multilingual native language identification. Natural Language Engineering, no. 1, pages 1–56.
12. *Shervin Malmasi* (2016), Native Language Identification: Explorations and Applications. Ph.D. thesis.
13. *Joel Tetreault, Daniel Blanchard, Aoife Cahill, Martin Chodorow* (2012), Native tongues, lost and found: Resources and empirical evaluations in native language identification. In Proceedings of the 24th International Conference on Computational Linguistics.
14. *Swanson, Eugene Charniak* (2012), Native Language Detection with Tree Substitution Grammars. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, vol. 2, pages 193–197.
15. *Marie-Catherine de Marneffe, Bill MacCartney, Christoper D. Manning* (2006), Generating Typed Dependency Parses from Phrase Structure Parses. In Proceedings of the Fifth International Conference on Language Resources and Evaluation.
16. *Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, Martin Chodorow* (2013), TOEFL11: A Corpus of Non-Native English. Technical report, Educational Testing Service.
17. *Scott Jarvis, Yves Bestgen, Steve Pepper* (2013), Maximizing classification accuracy in native language identification. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications. Atlanta, Georgia, pages 111–118.
18. *Andre Lynum* (2013), Native language identification using large scale lexical features. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 266–269.
19. *Marius Popescu, Radu Tudor Ionescu* (2013), The story of the characters, the dna and the native language. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 270–278.
20. *John Henderson, Guido Zarrella, Craig Pfeifer, John D. Burger* (2013), Discriminating non-native english with 350 words. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 101–110.

21. *Serhiy Bykh, Sowmya Vajjala, Julia Krivanek, Detmar Meurers* (2013), Combining shallow and linguistically motivated features in native language identification. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 197–206.
22. *Cyril Goutte, Serge Leger, Marine Carpuat* (2013), Feature space selection and combination for native language identification. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 96–100.
23. *Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, Tom Heskes* (2013), Improving native language identification with tf-idf weighting. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 216–223.
24. *Tomoya Mizumoto, Yuta Hayashibe, Keisuke Sakaguchi, Mamoru Komachi, Yuji Matsumoto* (2013), Naist at the nli 2013 shared task. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 134–139.
25. *Ching-Yi Wu, Po-Hsiang Lai, Yang Liu, Vincent Ng* (2013), Simple yet powerful native language identification on toefl11. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 134–139.
26. *Andrea Cimino, Felice Dell'Orletta* (2017), Stacked Sentence-Document Classifier Approach for Improving Native Language Identification. In Proceedings of the 12th Workshop on Building Educational Applications Using NLP.
27. *Ilia Markov, Lingzhen Chen, Carlo Strapparava, Grigori Sidorov* (2017), CIC-FBK Approach to Native Language Identification. In Proceedings of the 12th Workshop on Building Educational Applications Using NLP.
28. *Cyril Goutte, Serge Leger* (2017), Exploring Optimal Voting in Native Language Identification. In Proceedings of the 12th Workshop on Building Educational Applications Using NLP.