National Research University Higher School of Economics
Nizhny Novgorod

Authorship Attribution in Russian with New High-Performing and Fully Interpretable Morpho-Syntactic Features

Elena Pimonova, Oleg Durandin,
Alexey Malafeev

AIST Conference, 17-19 July 2019

Authorship Attribution

What do we solve?

- The task of identifying the author of a given text.
- The problem of modeling author’s style.

Why is this research relevant?

- There are not so many algorithms for Russian in comparison with English.
- Most existing methods don’t tell us anything about what author style is (although they show quite a high result in clustering and classification).

What is our goal?

- To increase the interpretability of text representation models in order to determine by which language means the author style is expressed.

Tools

- **SpaCy** library (https://spacy.io/) as convenient NLP pipeline (word and sentence tokenizer, morpho-syntactic analysis, etc.)
- **Russian language model** for spaCy (https://github.com/buriy/spacy-ru)
- **PyMorphy2** – Morphological analyzer/inflection engine for Russian/Ukrainian languages

Dataset

- **215 works of Russian literature (divided into blocks of 350 sentences = 1506 texts)**
- **30 authors**
- **18-21 centuries**

The material compiles with the following requirements:

- The selected authors are recognized by the international community (their works are presented in at least 5 world widest libraries).

- The selected authors are the «authors of the first row», that is, authors who introduced some changes to Russian literature.

- The selected works cover only one approximate period of the writer’s creative life.

Text Representation Models

- Simple Morphology and Syntax
- Complex Morphology and Syntax
- Treelet Bigrams and Trigrams
- Doc2Vec

Simple Morphology and Syntax Models

Simple Morphology Model

- relative frequencies for parts of speech in the text (e.g. NOUN, VERB, ADJ, etc.)
- 17 features

Simple Syntax Model

- relative frequencies for syntactic relations in the text (e.g. obj for direct object, etc.)
- 35 features

Complex Morphology Model

- new criteria for morphological markup
- word classification according to their semantic features (13 groups, e.g. *attribute*, *process*, etc.)

16 criteria for lexico-morphological analysis

- **Abstractness**
- **Pronominal replacement**
- **Action feature**
- **Generalized feature**
- **Descriptiveness**
- **Action descriptiveness**
- **Number**
- **Dynamism**
- **State**
- **Real modality**
- **Passive**
- **Present tense**
- **Past tense**
- **Future tense**
- **Action completeness**

- **E.g. Objectivity = (concrete nouns + pronouns) / content words**

Complex Syntax Model

- new criteria for syntactic markup
- **28 features** on two levels

Phrase level	Sentence level
Communication type (coordination, agreement, regimen, contiguity)	Contracted and uncontracted sentences
Structural type (complex phrase, simple phrase)	One-member and two-member sentences
Degree of phrase components unity (syntactically free and non-free phrase)	A number of complex structures (epenthetic construction, interjections, appeals, etc.)
Lexico-grammatical type (nominal phrase, verbal phrase, adverbial phrase)	

Treelet Bigrams and Trigrams

- Idea is taken from «Cross-lingual syntactic variation over age and gender» (Johannsen et. al)
- Treelets are typed relationships between tokens.

Bigram treelets

- dependency between main and dependent word:
  VERB → nsubj → NOUN

Trigram treelets

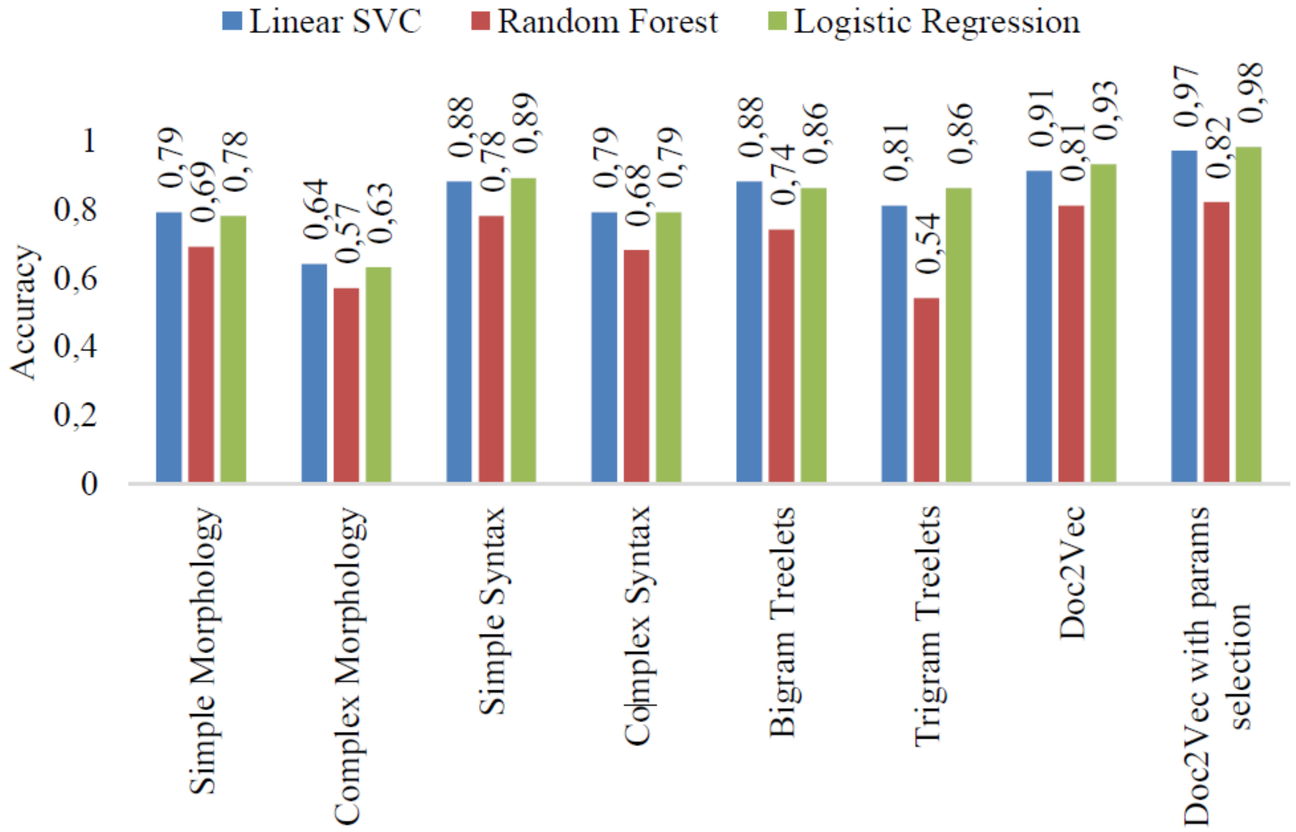
- two dependent words and one main word:
  NOUN ← VERB → NOUN
- consecutive subordination of words:
  VERB → NOUN → PRON

Doc2Vec

- Embedding technique
- Linking of words to each other in context
- Identifying the set of semantically close words for each author

Experiments

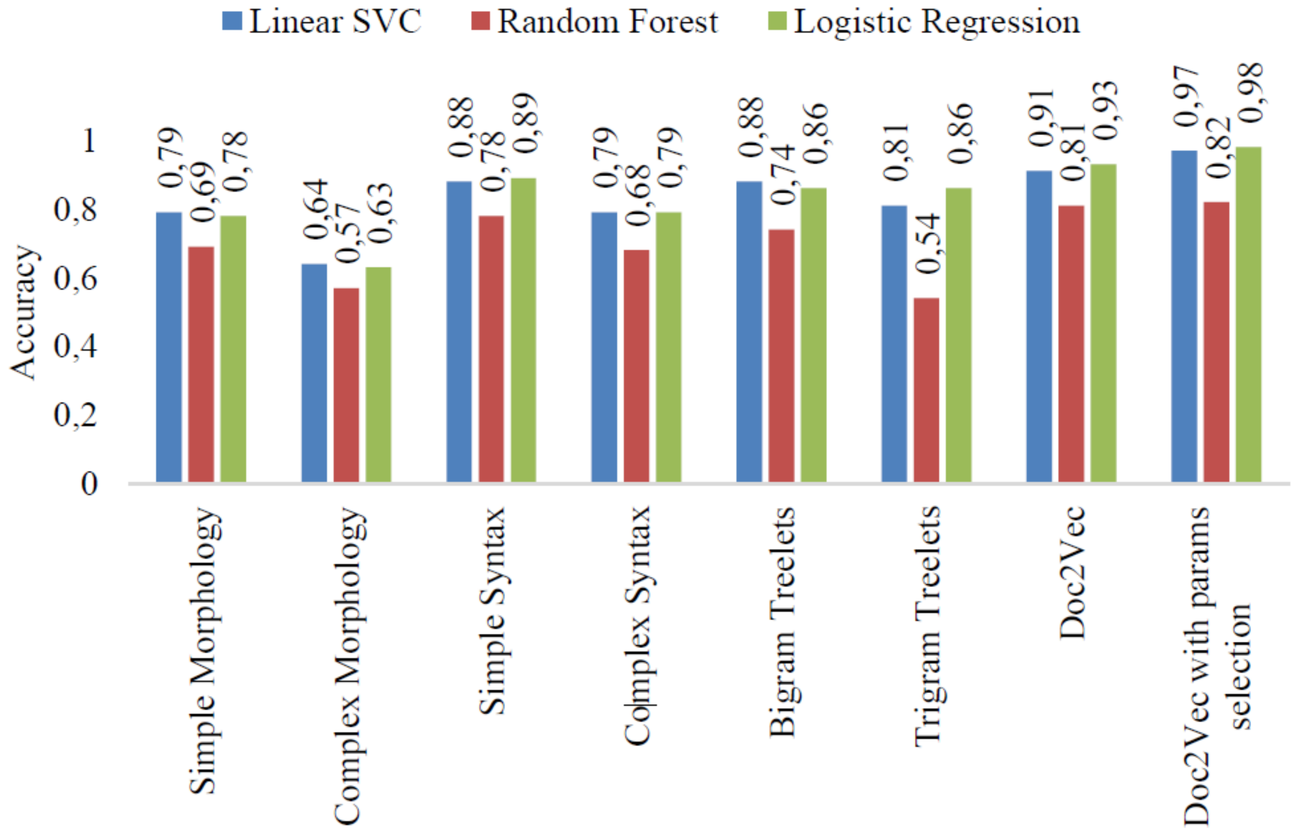
- Task of **multiclass classification (30 authors)**:
- Random Forest (20 base estimators);
- L_1-Logistic Regression (One-VS-Rest multiclassification type);
- SVM with a linear kernel;



Experiments

First conclusions

- Syntax-based models are more relevant for solving the authorship attribution problem than morphological ones.
- Simpler models consistently show better results than complex ones.



Experiments. Combination of Features

- **Combination led to increased classification accuracy.**
- Combination of all morphological and syntactic models showed result 94%.
- Their combination with the doc2vec model resulted in the highest accuracy 99%.

Experiments. Combination of Features

- The standalone use of morpho-syntactic features leads to quite good accuracy which proves their effectiveness for authorship attribution task.
- **Most importantly, they have the property of interpretability**.

Important Feature Analysis

Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
particle	–	discourse (emotional evaluation components)	–
conjunction	–	conj (relationships between homogeneous members)	homogeneous members as a complicator of the sentence
noun	objectivity (used in the text to state facts)	nsubj (connection between subject and predicate)	coordination and agreement
adverb	action feature and action descriptiveness	admod and advcl (relationship between the main word and modifier)	contiguity

Elements and relations at a simple level are part of a more complex level and continue to be assessed as important.

Error Analysis

- Confusion matrices analysis in all text representation models
- Styles of the authors who cannot be distinguished from each other may be similar.

1 group (0-3 errors): Sholokhov, Andreev, Gorky, Karamzin, Solzhenitsyn, Tolstoy, etc.
2 group (4-6 errors): Nabokov, Chernyshevsky, Goncharov, Lukyanenko, etc.
3 group (7+ errors): Vasilyev, Pushkin, Prishvin, Nosov, Gogol, Bulgakov.

- Some authors regularly had errors in different models of text representation.
- E.g. Bulychev and Nosov

Conclusion

- We used various text representation models in solving authorship attribution task.
- The best single model turned out to be the doc2vec with Logistic Regression (98%).
- Morpho-syntactic text representation models' standalone use yielded a comparable result (94%).
- Their combination with doc2vec improved the quality (99%).
- Proposed features are fully interpretable which makes it possible to determine linguistic markers of author's style.

Future Work

- stylometry (e.g. author profiling)
- plagiarism detection tasks
- cross-lingual aspect and identification of universal markers of style
- testing scalability of proposed approach

Code available: https://github.com/OlegDurandin/AuthorStyle

References

1. **Baayen, R., Halteren, H. van, Tweedie, F.**: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. Literary and Linguistic Computing 11(3), 121–132 (1996).

2. **Borisov, L., Orlov, Y., Osminin, K.**: Authorship attribution by the distribution of letter combination frequencies. 27th edn. Institute of Applied Mathematics named after M. Keldysh of the Russian Academy of Sciences, Moscow (2013).

3. **Dyachenko, P., Yomdin, L., Lasursky, A., Mityushin, L., Podlesskaja, O., Sizov, V., Frolova, T., Tsinman, L.**: The current state of the deeply annotated corpus of Russian language texts (SinTagRus). Proceedings of the Institute of Russian Language named after V.V. Vinogradov 6, 272–299 (2015).

4. **Johannsen, A., Hovy, D., Søgaard, A.**: Cross-lingual syntactic variation over age and gender. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning: CoNLL, pp. 103–112. Association for Computational Linguistics, Beijing (2015).

5**. Khmelev, D.**: Recognition of the text author using the Markov chains. MSU Bulletin 9 (2), 115–126 (2000).

6. **Korobov, M.**: Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Khachay M., Konstantinova N., Panchenko A., Ignatov D., Labunets V. (eds) Analysis of Images, Social Networks and Texts. AIST 2015. Communications in Computer and Information Science, vol 542, pp. 320–332. Springer, Cham (2015).

7. **Le, Q., Mikolov, T.**: Distributed representations of sentences and documents. In: ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning, pp. 1188–1196. JMLR, Beijing (2014)

8. **Luyckx, K., Daelemans, W., Vanhoutte, E.**: Stylogenetics: Clustering-based stylistic analysis of literary corpora. In: Proceedings of LREC-2006: The 5th International Language Resources and Evaluation Conference, Workshop Towards Computational Models of Literary Analysis, pp. 30–35. ILC, Genova (2006).

References

9. **Marneffe, M.-C. de, Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.**: Universal Stanford Dependencies: A cross-linguistic typology. In: Proceedings of the 9Th International Conference on Language Resources and Evaluation (LREC), pp. 4585–4592. European Language Resources Association (ELRA), Reykjavik (2014).

10. **OpenCorpora**, http://opencorpora.org/, last access 2019/04/30.

11. **Panicheva, P., Ledovaya, Y., Bogolyubova, O.**: Lexical, Morphological and Semantic Correlates of the Dark Triad Personality Traits in Russian Facebook Texts. In: Proceedings of the AINL FRUCT 2016 Conference, pp. 72–79. Institute of Electrical and Electronics Engineers Inc., St. Petersburg (2017).

12. **Pedregosa, F. et al.**: Scikit-learn: Machine Learning in Python. JMLR 12, 2825–2830 (2011).

13. **Poddubny, V., Shevelev, O., Kravtsova, A., Fatykhov, A.**: Vocabulary and analytical block of the Style Analyzer. In: 14th Russian Scientific and Practical Conference, pp. 138–140. Tomsk University, Tomsk (2010).

14. **Rogov, A., Sidorov, U., Solopova, A., Surovtsova, T.**: The information-analytical system “SMALT”. In: International Conference “Dialogue 2007”, pp. 470–474. Petrozavodsk State University, Bekasovo (2007).

15. **Russian language models for spaCy**, https://github.com/buriy/spacy-ru, last access 2019/04/21.

16. **Shvedova, N.**: Russian semantic dictionary. Explanatory dictionary, systematized by classes of words and meanings. 3d edn. Azbukovnik, Moscow (2003).

17**. spaCy**, https://spacy.io/, last access 2019/04/21.

NATIONAL RESEARCH
UNIVERSITY