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

Moscow, June 17–20, 2020

DOC2VEC OR BETTER INTERPRETABILITY? A METHOD STUDY FOR AUTHORSHIP ATTRIBUTION

Pimonova E. (hpimonova@gmail.com),
Durandin O. (oleg.durandin@gmail.com),
Malafeev A. (amalafeev@yandex.ru)

National Research University Higher School of Economics,
Nizhny Novgorod, Russia

In this work, we perform a method study for the problem of authorship attribution in Russian and English. The datasets used consist of 324 works written in Russian and 207 works in English. We propose a set of text representation models that reflect various linguistic phenomena, in particular, morphological and syntactic ones. One distinctive feature of the proposed models is that they are interpretable. These models are used individually and in combination against a Doc2Vec baseline. For Russian, some of our models outperform Doc2Vec, but this does not happen in the case of English, for various reasons. However, the proposed models can also be used together with Doc2Vec, dramatically improving its performance: by 16.79% in the case of Russian and by 7.2% for English. Additionally, we experiment with two different methods for separating texts into blocks of *K* sentences (contiguous and bootstrapped) and performed parameter tuning of *K*. Finally, we conduct a feature importance analysis and show which linguistic markers of author style are the most pertinent for Russian, English and for both these languages. All code used in this work is made freely available to the community[1].

Key words: authorship attribution, author style, text classification, text representation, feature engineering, natural language processing

DOI: 10.28995/2075-7182-2020-19-606-614

[1] https://github.com/OlegDurandin/AuthorStyle

DOC2VEC ИЛИ ИНТЕРПРЕТИРУЕМОСТЬ? СРАВНЕНИЕ МЕТОДОВ ОПРЕДЕЛЕНИЯ АВТОРСТВА

Пимонова Е. (hpimonova@gmail.com),
Дурандин О. (oleg.durandin@gmail.com),
Малафеев А. (amalafeev@yandex.ru)

1. Introduction

Authorship attribution is the task of determining the author of a written text based on a set of texts by candidate authors. Automatic algorithms for authorship attribution greatly simplify the solution of these problems and provide reliable and replicable results, which is especially important in criminal law and security matters. Most modern author attribution algorithms are based on formal and statistical models. Despite showing high accuracy in the classification problem, the algorithm results are difficult to interpret. In order to solve this problem, we propose more linguistically-grounded models for solving the attribution problem. We believe that our approach helps identify stylistic markers that can be used as guidelines when attributing a text to a particular author.

This paper is structured as follows. In **Section 2**, a literature review is given on the problem of authorship attribution. **Section 3** describes the proposed four text representation models. We report on the experiments conducted, discuss the results obtained and illustrate the findings of feature importance analysis in **Section 4**. Finally, **Section 5** gives some conclusive remarks.

2. Related Work

Approaches to authorship attribution can be formal or linguistics-based. One of the most common formal models is the n-gram model. Some of the first published algorithms for authorship attribution in Russian used character bigrams [Khmelev, 2000] and trigrams [Borisov, et al. 2013]. N-grams are successfully used for author profiling as well; in the work by Litvinova et al. [2018] on age identification using the first age-annotated corpus for the Russian language, the authors supplemented word n-grams with part-of-speech n-grams.

In the PAN competitions [Kestemont et al. 2019], the baseline character trigram model was improved on by using variable-length character and word n-grams [Custodio and Paraboni 2018], as well as by extracting n-grams after text distortion [Muttenthaler, et al. 2019]. Other n-gram-based models [Murauer et al., 2018]; [Bacciu et al., 2019] also showed high accuracy in the PAN competitions.

Other formal approaches to solving the attribution problem are text compression [Halvani and Graner 2019] and frequency analysis of various text features: word

frequencies [Poddubny, et al. 2010], the number of sentences, text length, character and punctuation frequencies [Safin and Ogaltsov 2018].

When linguistic (and usually language-dependent) features are used, morphology and syntax are most commonly modeled [Baayen et al. 1996], [Rogov et al. 2007], [Hosseinia and Mukherjee 2018]. Other linguistic methods involve modeling semantics [Panicheva et al. 2016].

The linguistic approach to authorship attribution is not as widespread as formal language-independent models, but it performs on par with them. While linguistic models are language-dependent, they are often more interpretable. In this work, we propose morphology and syntax models for the Russian and English languages. We believe that these models can help identify reliable stylistic markers that are useful both for computational analysis of author style and for text analysis performed by human experts.

3. Text Representation Models

3.1. Doc2Vec

We used five text representations. The Doc2Vec [Le, Mikolov 2014] model was chosen as the baseline. It is an embedding model for representing sentences, paragraphs or entire documents as vectors. Doc2Vec is known to perform well on various text classification tasks. To improve the quality of the baseline model, we developed two morphological and two syntactic models that differ in representation complexity.

3.2. Simple Morphology and Syntax Models

The so-called 'simple' morphological and syntactic models include relative frequencies of parts of speech and syntactic relations present in the text. The number of the morphological features (17, including punctuation, special characters, and foreign words) is the same for Russian and English since we used the language-agnostic UDPipe tool [Straka et al. 2016]. In the simple syntax model, we identified 38 types of syntactic relations such as *nsubj* (subject) and *fixed* (non-free phrase) for Russian and 45 for English.

3.3. Complex Morphology Model

To increase interpretability, we developed the so-called 'complex' morphology and syntax models that encompass higher-level language phenomena. The complex morphological model relies on semantic features of words (e.g. the noun "running" denotes a process, etc.). In this model, we used the OpenCorpora markup for the Russian language, since it distinguishes between a larger number of morphological types than Universal Dependencies. The English model still used the UD markup, which resulted in a loss of some features available for Russian (16 versus 10 features).

The semantic attributes used are closely tied to morphological characteristics of words, hence the name of the model. The most ambiguous part of speech in terms of determining the semantic attribute was the noun. We used the "Russian semantic

dictionary" by N. Yu. Shvedova for grouping nouns based on their semantic features. An example of a feature under this model is "dynamism", or the ratio of words with the semantic attribute *process* to all content words. This criterion allows one to determine how much the author is inclined to narration and active change of action.

There were also criteria in our model that took into account the morphological characteristics of the entire text. Some examples are the proportion of verbs in the passive voice or verbs in the past tense to all verbs.

3.4. Complex Syntax Model

Similarly, a complex syntax model was developed, with distinguishing features at the phrase and sentence level. Phrases are categorized according to communication type (coordination, agreement, verb government, or contiguity), structural type (simple and complex phrase), the degree of phrase component unity (syntactically free and non-free phrases) and lexical-grammatical type (nominal, verbal and adverbial). Each criterion is represented by several types of relations, normalized by the total number of relations. For example, the proportion of syntactically non-free phrases in the text was calculated by the formula: (*flat* + *fixed* + *compound*) / *N*, where *flat* is the number of named entities, *fixed* is non-free phrases, *compound* is compound and composite numerals, N is the total number of syntactic relations.

At the sentence level, we considered contracted sentences, vocatives, genitives, various types of one-member sentences and semi-complex sentences. These parameters were calculated taking into account not only the syntactic relations representing the class, but also the morphological characteristics of the words associated with these syntactic relations. For example, indefinite-personal sentences include those that do not have the relations *nsubj* and *csubj* (the connection between the subject and the predicate) coming from the root word. In this case, the root word must also be a verb either in 3rd person plural form, present or future tense, e.g. *govoryat* (≈*people say*) or in the form of the plural past tense, like *pogovarivali* (≈*there were rumors*). Adapting the originally Russian-based complex syntax model to the English language, we omitted genitive sentences and one-member sentences (except nominative ones), since in English most well-formed sentences have a subject. Thus, we got 28 features for Russian and 22 for English.

4. Experiments

4.1. NLP Framework and Dataset

As mentioned above, we relied on the UDPipe library as the natural language processing framework. The following language models were used: English-EWT and Russian-SyntagRus.

For Russian, we used a corpus that contains 324 works of Russian literature, created by 30 authors spanning XVIII–XXI centuries. For English, we selected 207 works by 34 classical English authors from the Gutenberg Project (gutenberg.org). We divided the entire set of works into training and test sets in such a way that all

authors were present, but different works by these authors were used for training and testing, like in PAN competitions. For Russian, the training set included 186 texts, ~5M words, while the test set had 138 texts, ~2.5M words. For English, there were 137 texts, ~15M words in training and 70 texts, ~7.4M words in the test dataset.

4.2. Evaluation Method

Following competitions such as PAN [Kestemont et al. 2019], we used classification accuracy as an evaluation metric, that is, the proportion of works whose authors were correctly attributed by the system. Since many literary works in the dataset are quite large, we divide them into blocks. Each of the blocks is classified by the system, then a prediction is made by majority vote as to who authored the entire text. Only final, post-vote predictions are evaluated.

4.3. Experiment Setting

In the experiments, the representations proposed in **Section 3** were evaluated against the Doc2Vec baseline, independently and in various combinations. Classic machine learning algorithms were used, namely logistic regression with L1 regularization, random forest, and a linear SVC.

Apart from text representations, we also experimented with some methods for separating texts into blocks of K sentences. The value K is a hyperparameter that affects classification accuracy, so we performed some parameter tuning. We tested two alternative approaches to extracting blocks of text: contiguous (non-overlapping blocks) and bootstrapped (blocks can overlap and are randomly sampled from each text).

4.4. Results and Discussion

We will only list results for logistic regression because it significantly outperformed random forest and linear SVC. Due to size constraints, we will not show the results obtained with each configuration (language, text representation or a combination of text representations, machine learning algorithm, the value of K and text sampling strategy) that we tested, of which there were over 500. Only the best results for each text representation model will be discussed in this section (see **Table 1** and **Table 2**). Optimized Doc2Vec parameter values were as follows: window = 10, min_count = 3, negative = 5, vector size = 100.

As can be observed for both Russian and English, the complex morphology and syntax models, when used individually, performed much worse than the simple morphology and syntax models, respectively. For Russian, the simple syntax model outperformed the baseline Doc2Vec method, while for English none of the proposed models (or their combinations) surpassed the baseline. This is partly due to the fact that the complex morphology and syntax models were originally developed for Russian, so they had to be somewhat simplified to accommodate English. Another factor in the higher accuracy of morphosyntactic features for Russian is that, unlike English, Russian is a morphologically-rich language and thus authors have more tools for expression at this level. For English, however, lexical features (as captured by Doc2Vec in our approach) are much more powerful.

Table 1. Authorship Attribution Classification Accuracy on the Russian-language test set of 138 texts
(Legend: SM—simple morphology, CM—complex morphology, SS—simple syntax, CS—complex syntax, SMS—simple morphology and syntax, CMS—complex morphology and syntax, SCMS—simple + complex morphology and syntax, K—number of sentences per block of text, CB—contiguous blocks, BB—bootstrapped blocks)

Configuration	Classification accuracy
SM, K=350, BB	0.511
CM, K=500, BB	0.430
SS, K=500, BB	0.693
CS, K=500, CB	0.526
SMS, K=500, BB	0.737
CMS, K=350, BB	0.693
SCMS, K=400, BB	0.774
Doc2Vec, K=300, CB	0.613
Doc2Vec + SMS, K=450, CB	0.766
Doc2Vec + SCMS, K=400, CB	**0.781**

Table 2. Authorship Attribution Classification Accuracy on the English-language test set of 70 texts

Configuration	Classification accuracy
SM, K=300, BB	0.6
CM, K=300, BB	0.371
SS, K=400, BB	0.773
CS, K=300, BB	0.586
SMS, K=400, CB	0.787
CMS, K=300, BB	0.671
SCMS, K=400, BB	0.792
Doc2Vec, K=400, CB	0.886
Doc2Vec + SMS, K=300, CB	0.929
Doc2Vec + SCMS, K=400, CB	**0.957**

Importantly, the proposed text representation models succeed in improving Doc2Vec results. In particular, the combination of all four proposed models (SCMS) resulted in an improvement of 16.79% over Doc2Vec for Russian. For English, the improvement was 7.2%, still very considerable.

5. Feature Importance Analysis

We conducted an analysis of important features in each of the four proposed text representation models to determine which linguistic markers help distinguish one author from another. Tables 3–5 list such style markers for both Russian and English, as well as language-specific ones.

Table 3. Style markers for both Russian and English

	Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
Universal	Function words (conjunctions and particles)	—	conj—relationship between homogeneous members, cc—connection with a means of communication	Homogeneous members
	Noun	—	nsubj—connection between subject and predicate	—
	Punctuation	—	—	Complex structures (participle, adjective and verb-adverb constructions)

Table 4. Russian-specific style markers

	Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
Russian	Adverb	Action feature, action descriptiveness (used in the text to describe an action)	advmod, advcl—connection with adjunct	Contiguity linkage
	Noun	**Abstractness** (used in the text to state abstract notions), **objectivity** (used in the text to state facts)	nsubj—connection between subject and predicate	**Coordination and agreement linkage**
	Pronoun	Pronominal replacement (used to replace a noun or noun phrase)	—	—

Table 5. English-specific style markers

	Simple Morphology	Complex Morphology	Simple Syntax	Complex Syntax
English	—	—	flat—relationship between named entities,	Syntactically non-free combinations
	Auxiliary words	—	aux—connection with an auxiliary word	—
	—	Real modality, passive voice	—	—

6. Conclusion and Future Work

Thus, we have proposed and tested a novel approach to authorship attribution that consists in supplementing Doc2Vec with frequencies of parts of speech and syntactic relations, as well as with manually-designed features that reflect larger-scale morphological and syntactic phenomena relevant to author style. This approach is suitable for Russian and English, although we found that lexis has a lot more impact on authorship attribution accuracy in English, while the proposed features at the morphology and syntax levels perform much better on Russian texts.

It is also worth noting that our approach is only suitable for larger chunks of text. The performance with $K < 250$ drops significantly.

For future work, it is possible to test similar approaches on other languages and perform a comparative study of feature importance for a larger set of languages.

References

1. *Baayen R., Halteren H. van, Tweedie F.* (1996), Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, Vol. 11(3), pp. 121–132.
2. *Bacciu A., Morgia M. La, Mei A., Nemmi E. N., Neri V., Stefa J.* (2019), Cross-Domain Authorship Attribution Combining Instance-Based and Profile-Based Features: Notebook for PAN at CLEF 2019, available at: http://ceur-ws.org/Vol-2380/paper_220.pdf.
3. *Borisov L. A., Orlov Y. N., Osminin K. P.* (2013), Authorship attribution by the distribution of letter combination frequencies [Identifikatsiya avtora teksta po raspredeleniyu chastot bukvosochetaniy], Keldysh Institute Preprints [Preprinty IPM im. M. V. Keldysha], Vol. 27, pp. 3–26.
4. *Custódio J. E., Paraboni I.* (2018), EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_76.pdf.
5. *Halvani O., Graner L.* (2018), Cross-Domain Authorship Attribution Based on Compression: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_90.pdf.
6. *Hosseinia M., Mukherjee A.* (2018), A Parallel Hierarchical Attention Network for Style Change Detection: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_91.pdf.
7. *Kestemont M., Stamatatos E., Manjavacas E., Daelemans W., Potthast M., Stein B.* (2019), Overview of the cross-domain authorship attribution task at {PAN} 2019, CEUR Workshop Proceedings, Vol. 2380, pp. 1–15.
8. *Khmelev D. V.* (2000), Recognition of the text author using the Markov chains [Raspoznavaniye avtora teksta s ispolzovaniyem tsepey A.A. Markova], MSU Bulletin [Vestnik MGU], Vol. 9 (2), pp. 115–126.
9. *Le Q., Mikolov T.* (2014), Distributed representations of sentences and documents, n Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pp. 1188–1196.

10. *Litvinova T. A., Sboev A. G., Panicheva P. V.* (2018), Profiling the Age of Russian Bloggers, Proceedings of the 7th International Conference, AINL 2018, St. Petersburg, pp. 167–177.
11. *Murauer B., Tschuggnall M., Specht G.* (2018), Dynamic Parameter Search for Cross-Domain Authorship Attribution: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_84.pdf.
12. *Muttenthaler L., Lucas G., Amann J.* (2019), Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams: Notebook for PAN at CLEF 2019, available at: http://ceur-ws.org/Vol-2380/paper_49.pdf.
13. *Panicheva P. V., Ledovaya Y. A., Bogolyubova O. N.* (2016), Lexical, Morphological and Semantic Correlates of the Dark Triad Personality Traits in Russian Facebook Texts, Proceedings of the AINL FRUCT 2016 Conference, St. Petersburg, pp. 72–79.
14. *Poddubny V. V., Shevelev O. G., Kravtsova A. S., Fatykhov A. A.* (2010), Vocabulary and analytical block of the Style Analyzer [Slovarno-analiticheskiy blok sistemy "Stileanalizator"], 14th Russian Scientific and Practical Conference [Nauchnoye tvorchestvo molodezhi: Materialy XIV Vserossiyskoy N76 nauchno-prakticheskoy konferentsii], Tomsk, pp. 138–140.
15. *Rogov A. A., Sidorov U. V., Solopova A. I., Surovtsova T. G.* (2007), The information-analytical system "SMALT" [Informatsionno-analiticheskaya sistema "SMALT"], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2007" [Kompyuternaya lingvistika i intellektualnyye tekhnologii: Trudy Mezhdunarodnoy konferentsii "Dialog 2007"], Bekasovo, pp. 470–474.
16. *Safin K., Ogaltsov A.* (2018), Detecting a Change of Style Using Text Statistics: Notebook for PAN at CLEF 2018, available at: http://ceur-ws.org/Vol-2125/paper_104.pdf.
17. *Shvedova N. Yu.* (2003), Russian semantic explanatory dictionary. Explanatory dictionary, systematized by classes of words and meanings [Russkiy semanticheskiy slovar. Tolkovyy slovar. sistematizirovannyy po klassam slov i znacheniy]. Azbukovnik, Moscow.
18. *Straka M., Hajic J., Straková J.* (2016), UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4290–4297.