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

Moscow, June 17–20, 2020

ADDRESSING THE AUTHORSHIP ATTRIBUTION PROBLEM FOR DIFFERENT SEGMENTS OF THE INTERNET

Solonin M. A. (mikhail.solonin@abbyy.com)
ABBYY, Moscow, Russia

The purpose of this study is to evaluate the possibilities of authorship attribution methods for different segments of the Internet. We focus our efforts on Russian-language content. Since texts in different segments of the Internet have various features and characteristics, we collected datasets from different sections, such as channels messengers, blogs, news reviews and literary works.

Key words: authorship attribution, text deanonymization

DOI: 10.28995/2075-7182-2020-19-1160-1169

РЕШЕНИЕ ЗАДАЧИ АВТОРСКОЙ АТРИБУЦИИ ДЛЯ РАЗЛИЧНЫХ СЕГМЕНТОВ ИНТЕРНЕТА

Солонин М. А. (mikhail.solonin@abbyy.com)
ABBYY, Москва, Россия

Главная задача этого исследования — оценить возможности методов автоматической авторской атрибуции для некоторых частей Интернета. Отметим, что приоритетными в данной работе являются тексты на русском языке. Так как тексты из различных сегментов Интернета могут иметь разные особенности и характеристики, мы собрали данные из следующих разделов: каналы в мессенджерах, персональные блоги, новостные обзоры, литературные работы.

Ключевые слова: авторская атрибуция, деанонимизация текста

1. Introduction

In today's world, Internet communication and online media are playing an increasingly important role. Their development leads to the fact that for many people Internet resources have become the main or the only source of news and relevant information. In such a context, we must have mechanisms to deal with the threats posed by the Internet. The ability of text deanonymization will reduce the threat of extremism, cyberbullying and the spread of fake news. In addition, the author identification problem as an independent task is of great importance from the philological and literary points of view. To evaluate the ability of some models, we have collected several datasets of Russian texts. We will discuss them in detail in the "Data description" section.

2. Problem setting and metrics

Let us first formalize the problem. Given a corpus with a fixed number of authors. We create and train the model. Further, for each input text whose authorship is unknown, we want to predict the author. We assume that we know *a priori* that the author of the input text was presented in the training set. To evaluate the quality of our models, we will use **F1-score** with macro average and **Accuracy** score.

3. Data description

As we need to explore different parts of Internet textual content and develop models for author identification, we collected data from the following sources:

1. **Meduza** news overview—analysis and description of the news from the "History" section. 29 authors.
   Key features: dry, factual statements, medium length texts.
2. **Telegram**'s political channels—reaction to the news agenda. 50 authors.
   Key features: short messages, abundance of profanity.
3. **VK** blogs—blogs of business coaches. 15 authors.
   Key features: medium length posts, no thematic bias between the texts in the dataset.
4. **Snob**.ru blogs—personal blogs. 15 authors.
   Key features: medium length posts, pronounced author's style.
5. **Zhurzal** dataset—part of The General Internet-Corpus of Russian (GICR). Contains fiction and nonfiction literature. 16 authors.
   Key features: long texts, potential thematic bias.

Figure 1 shows the distribution of texts lengths and some statistical data. Mean and median length values for each dataset are marked with brown and olive, respectively.

The graphs highlight the differences in statistical values for texts from different sources. Examples of texts can be found in the appendix.

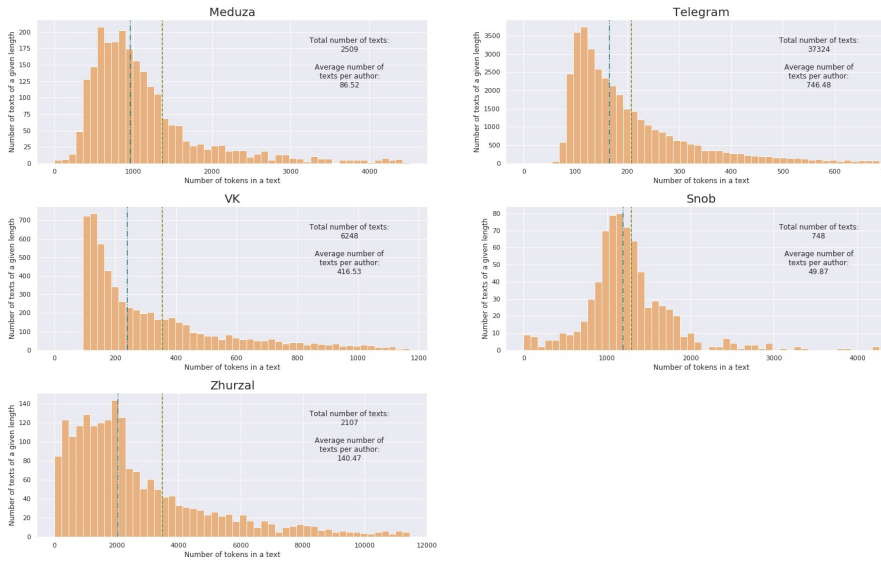

Figure 1: Texts length distribution

4. Methods

We developed and tested several models to solve the problem of authorship attribution:

1. tf-idf + classifier
   Here we used XGBoost classifier over tf-idf vectorized texts.
2. BiLSTM + Attention
   The model described in ("Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification", Peng Zhou et al)
3. CharCNN+BiLSTM+Pooling
   Here and below it is assumed that token and character representations are trainable.
   We take token embeddings from GloVe.
4. CharCNN+Bert+BiLSTM+Pooling
   Here and below it is assumed that BERT layers are frozen.
   We take pretrained BERT for Russian from DeepPavlov (RuBERT).
5. CharCNN+Bert+BiLSTM+Multi-features-pooling
6. Bert classifier fine tuning
   Based on ("How to Fine-Tune BERT for Text Classification?", Chi Sun et al)

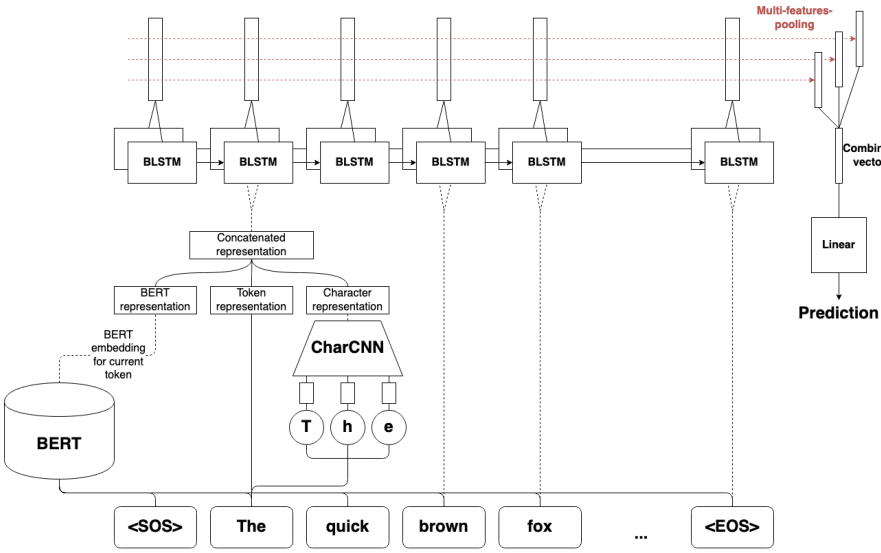

Figure 2: CharCNN+Bert+BiLSTM+Multi-features-pooling scheme

5. Models evaluation

The first thing we notice is that (6) Bert fine tuning showed very poor performance for this problem, and therefore was excluded from the considered models.

When evaluating models, (5) CharCNN+Bert+BiLSTM+MF-pooling proved to be the most efficient and robust method. Therefore, further experiments were performed with this model. Table 1 contains metric values for different datasets with the following setting:

- BERT tokenizer
- no preprocessing
- original datasets (without augmentation)

Table 1: Models evaluation

Dataset	Method	F1-score macro average	Accuracy
MEDUZA	TF-IDF + classifier	0.32	0.49
	CharCNN+BLSTM+Pooling	0.36	0.43
	CharCNN+BLSTM+BERT+Pooling	0.41	0.51
	CharCNN+BLSTM+BERT+MF-Pooling	**0.44**	**0.57**
Telegram	TF-IDF + classifier	0.22	0.41
	BLSTM + Attention	0.207	0.317
	CharCNN+BLSTM+Pooling	0.4	0.53
	CharCNN+BLSTM+BERT+Pooling	0.42	0.61
	CharCNN+BLSTM+BERT+MF-Pooling	**0.42**	**0.61**

Dataset	Method	F1-score macro average	Accuracy
VK	TF-IDF + classifier	0.25	0.64
	BLSTM + Attention	0.37	0.55
	CharCNN+BLSTM+Pooling	0.64	0.84
	CharCNN+BLSTM+BERT+Pooling	0.78	0.82
	CharCNN+BLSTM+BERT+MF-Pooling	**0.78**	**0.82**
SNOB	TF-IDF + classifier	0.52	0.72
	CharCNN+BLSTM+Pooling	0.64	0.78
	CharCNN+BLSTM+BERT+Pooling	0.82	0.87
	CharCNN+BLSTM+BERT+MF-Pooling	**0.82**	**0.87**
Zhurzal	**TF-IDF + classifier**	**0.89**	—
	BLSTM + Attention	0.64	—
	CharCNN+BLSTM+BERT+Pooling	0.35	—

6. Experiments

6.1. Different tokenization types

We used different tokenizers:

- nltk-style tokenizer
- BERT tokenizer (i.e. BPE-tokenization)
- space symbols tokenization

and different text preprocessing methods:

- special symbols removal
- conversion of special symbols to standard (e.g. quotation marks cast)
- lowercase cast
- no preprocessing

Our experiments showed that the choice of a tokenization method or preprocessing does not significantly affect the performance of neural models.

This can be explained by the fact that neural network methods are very flexible for the above transformations and still can capture the key features of sentences.

6.2. Removal of de-anonymization features

For the VK dataset we conducted the following experiment:

1. for each author we found several most common entities (phrases, smiles, etc.)
2. we emit 2 datasets: the first one contains text without the potentially de-anonymization entities, in the second one they are removed
3. we train 2 models and evaluate them

This experiment showed, that some introducing and concluding phrases, that are mostly used by one certain author, may be considered to be de-anonymization entities. However, smiles and single words, used frequently by one author and rarely by other, do not tend to identify the authorship.

6.3. Removal of authors with a small number of texts

For the Meduza dataset we conducted the following experiment:

1. we choose some model and train it on a full dataset
2. we exclude from the dataset texts of 6 authors, that have the lowest number of texts
3. we train a new model and compare it's predictions quality with the quality of the original model and dataset

In such a setting, we expect to have the same performance. However, all neural networks except (5) CharCNN+BLSTM+BERT+MF-Pooling had significantly higher quality on the "truncated" dataset. This means that only (5) CharCNN+BLSTM+BERT+MF-Pooling is robust and has no bottleneck in it's architecture.

6.4. Lemmatization and permutation of words in the text

1. We tried to augment datasets with lemmatized versions of the texts, but it led to decline of the model performance.
2. We tried to augment datasets by swapping some words in the text. Permutation of words was made according to the following scheme: First we select p_1 % of the tokens from the text, then change them with the token having the position $current\_token\_position + normal(\mu, \sigma)$. We tested this method for $p_1 \in \{10, 15\}$, $\mu = 0$ and $\sigma \in \{3, 5\}$. This method provided little quality growth only for VK.

6.5. BiLSTM + Multi-head Attention

For model (2) BiLSTM + Attention, we tried to use a basic idea, but with multi-headed attention. However, this approach did not lead to an increase in quality on the VK dataset. Moreover, this method was not robust for a Meduza dataset.

7. Results analysis

The evaluation of the models shows that each dataset has it's own features and characteristics. Let us give a brief explanation of these results.

The Zhurzal dataset consists of texts of large size and on various subjects, which is ideal for the good work of tf-idf. At the same time, neural network methods work worse because they cannot fully use the entire text for analysis. Also, the texts from Zhurzal are heterogeneous on the subject (unlike all other cases), which also gives tf-idf advantage.

For all other cases neural network methods proved to be better than tf-idf (the best result is achieved using BERT embeddings, which is expected).

The correlation of the results with the corpora seems understandable: Medusa has dry, factual, maximally depersonalized texts, and the quality of the classifiers on it is the weakest. Telegram consists of short texts, and also has the largest number of classes, so the quality on it is also not high. Snob and VK seem to be the most interesting setting: the texts are long enough to catch the author's style, but at the same time, the problem does not become the task of thematic modeling.

8. Further work

In future works, we would like to explore some other Internet resources, as well as focus on creating more subtle methods. To solve authorship attribution problem on texts of short length or when having a small amount of data, it is important to capture the stylistic or linguistic features of the texts for each author. We expect to conduct research with a deeper understanding of the individual characteristics of the authors.

9. Conclusion

In this study, we analyzed various segments of the Russian-language Internet and evaluated the possibilities of authorship identification methods. Analysis of the results shows that each section of the Internet has its own characteristics and features. Although neural methods show relatively good quality on medium-length texts, we still have to work on improving methods for short and long messages.

Appendix

1. Meduza
   мэрия москвы заказала проекты новых домов для реновации. они будут выглядеть примерно так московский фонд реновации жилой застройки , который будет отвечать за массовое переселение москвичей из пятиэтажек , объявил конкурсы на проекты первых 20 домов (по трем округам—остальные объявят позже) . ожидается , что итоги тендеров подведут в конце января , а сами проекты домов будут готовы летом 2018 года , после чего начнется строительство . <…>
   — Иван Голунов

2. VK
   никогда не нужно опускать руки . несмотря на то , что интеллектуальные возможности определяются в возрасте до 5 лет—это не повод прекращать свое развитие во взрослой жизни . человек в детстве получает 80 % информации , но у вас всегда остаются резервные 20 % . главное—продолжать действовать . люди , которые добиваются посредственных результатов , почти не отличаются от тех , кто добивается выдающихся

результатов . но те , кто добивается выдающихся результатов , просто что—то делают немного иначе . их волевой потенциал лишь чуть лучше , но именно он является определяющим . <…>
— Радислав Гандапас

3. Snob
тадеуш мазовецкий создал современную польшу . авторитарный режим был расформирован постепенно , в ходе переговоров с оппозицией . а после , без особого драматизма , страна стала жить по новым , вполне демократическим правилам . таков главный урок 1989 года—года мазовецкого . это стало возможно не сразу . вначале « солидарность » едва не привела коммунистическое государство в состояние полной недееспособности . потом было введено военное положение . наконец в польских верхах возникло понимание , что контроль—синоним слова « тупик » , а из тупика без помощи оппозиции не выбраться . <…>
— Станислав Кувалдин

почему в творчестве шестидесятников « течет шампанское рекою » в 1966 году моей курсовой работой о неологизмах велимира хлебникова заинтересовалось не только кгб . « здравствуйте , это поэт евтушенко » ,—раздалось в трубке телефона летом после первого курса . оказалось , что на фестивале негритянского искусства литераторы евтушенко и долматовский познакомились с моим отцом—послом в сенегале . родители рассказали об увлечении сына хлебниковым—уникальном для того времени интересе . <…>
— Виктор Ерофеев

4. Telegram
так глядишь поменяют конституцию и прямые выборы канцлера введут в германии . . . . президент германии франк—вальтер штайнмайер в своем первом рождественском обращении к гражданам фрг призвал их не бояться затянувшегося процесса формирования правительства и доверять властям . . . . вот так медленно уничтожают репутацию меркель : нет решения ни по коалиции , ни решения о досрочных выборах , что ставит вопрос : а зачем тогда вообще нужны такие выборы и такое правительство , если после них правят всё те же министры только в статусе врио ?
— Политджойстик

5. Zhurzal
кому , как не григорию кружкову , переводчику стольких замечательных английских стихов , говорить о связи русской и англоязычных литератур ? перевод , кроме всего прочего , требует эрудиции и способности к долгой и упорной работе книга еще раз показывает , что этими качествами кружков обладает вполне . он проделывает огромную работу по разысканию упоминаний классика ирландской поэзии йейтса в русской периодике и переписке литераторов начала xx века и проводит

параллели между йейтсом и русским поэтом мифотворцем вячеславом ивановым в образах и сюжетах . даже башни были у обоих . <...>
—**Александр Уланов**

марк далет . орбинавты роман . м . новое литературное обозрение , 2011 . по совести сказать , то , что появление орбинавтов марка далета с самого начала , еще до выхода книги , сопровождалось активной рекламной кампанией , вызывало у меня сильный скепсис . книга еще не вышла , а из нее уже публиковались отрывки ; едва обрела бумажную плоть , а у нее уже был целый собственный сайт и даже буктрейлер , невиданный по крайней мере , не слишком еще привычный зверь в наших широтах (видеоролик о книге , призванный убеждать читателей в том , как все это интересно) . <...>
— **Ольга Балла**

блистательный князь как то по вечернему петербургу шли двое . один в простом военном мундире , другой в щегольском кафтане . настроение у попутчиков было весёлое , они травили анекдоты , как вдруг тот , что в мундире , явственно различил голос павел , бедный павел , бедный князь ! он невольно вздрогнул , остановился и оглянулся . перед глазами предстал таинственный некто в испанском плаще , со шляпой , надвинутой на глаза . сомнений быть не могло орлиный взор , смуглый лоб и строгая улыбка выдавали великого прадеда павла петра i .
—**Лев Бердников**

References

1. *Haifa Alharthi, Diana Inkpen, and Stan Szpakowicz.* Authorship identification for literary book recommendations. In Proceedings of the 27th International Conference on Computational Linguistics, pages 390–400, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.
2. *Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.* Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.
3. *Hugging Face Tokenizers framework.* https://github.com/huggingface/tokenizers. 20-02-2020.
4. *Yurii Kuratov and Mikhail Arkhipov.* Adaptation of deep bidirectional multilingual transformers for russian language. 05 2019.
5. *Multi label Text Classification using BERT.* https://github.com/kaushaltrivedi/fast-bert. 20-02-2020.
6. *Sreenivas Mekala, Vishnu Vardan Bulusu, and T. RaghunadhaReddy.* A survey on authorship attribution approaches. 2018.
7. *The General Internet-Corpus of Russian (GICR).* http://www.webcorpora.ru/en/. 20-02-2020.
8. *Jagadeesh Patchala and Raj Bhatnagar.* Authorship attribution by consensus among multiple features. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2766–2777, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

9. *Jeffrey Pennington, Richard Socher, and Christopher D. Manning.* Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.
10. *Kröll-M., Ziak H., Rexha, A. et al.* Authorship identification of documents with high content similarity. page 223–237. Scientometrics 115, 2018.
11. *Liuyu Zhou and Huafei Wang.* News authorship identification with deep learning. 2016.
12. *Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu.* Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 207–212, Berlin, Germany, August 2016. Association for Computational Linguistics.