2023

Impact of Character n-grams Attention Scores for English and Russian News Articles Authorship Attribution

Liliya Mukhmutova
Technological University Dublin, Ireland, liliya.makhmutova@tudublin.ie

Robert J. Ross
Technological University Dublin, Ireland, robert.ross@tudublin.ie

Giancarlo Salton
Universidade comunitaria da Regiao de Chapeco, Santa Catarina

Recommended Citation

Mukhmutova, Liliya; Ross, Robert J.; and Salton, Giancarlo, "Impact of Character n-grams Attention Scores for English and Russian News Articles Authorship Attribution" (2023). *Conference papers*. 13.
https://arrow.tudublin.ie/diraacon/13

Funder: SFI through ML-Labs 18/CRT/6183

Impact of Character n-grams Attention Scores for English and Russian News Articles Authorship Attribution

Liliya Makhmutova
Technological University Dublin
Dublin
d21124385@mytudublin.ie

Robert Ross
Technological University Dublin
Dublin
robert.ross@tudublin.ie

Giancarlo Salton
Unochapecó - Universidade
Comunitária da Região de Chapecó
Chapeco, Santa Catarina
gian.salton@gmail.com

ABSTRACT

Language embeddings are often used as black-box word-level tools that provide powerful language analysis across many tasks, but yet for many tasks such as Authorship Attribution access to feature level information on character n-grams can provide insights to help with model refinement and development. In this paper we investigate and evaluate the importance of character n-grams within an embeddings context in authorship attribution through the use of attention scores. We perform this investigation both for English (Reuters_50_50) and Russian (Taiga) news authorship datasets. Our analysis show that character n-grams attention score is higher for n-grams that are considered to be important for authorship identification for humans. Beyond specific benefits in authorship attribution, this work provides insights into the importance of character n-grams as a unit within embeddings.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**;

KEYWORDS

Character n-grams, Authorship Attribution task, attention score

ACM Reference Format:
Liliya Makhmutova, Robert Ross, and Giancarlo Salton. 2023. Impact of Character n-grams Attention Scores for English and Russian News Articles Authorship Attribution. In *The 38th ACM/SIGAPP Symposium on Applied Computing (SAC '23), March 27-March 31, 2023, Tallinn, Estonia.* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.1145/3555776.3577856

1 INTRODUCTION

In the past 10 years word embeddings based techniques have become dominant across a range of Natural Language Processing (NLP) tasks, such as language modelling, machine translation, text generation, text classification, and intent recognition etc. Advances such as the introduction of transfer learning to NLP in 2018 [1] and the advent of attention models and then the transformer architecture [2], have lead to a wide variety of embeddings models which can be customised to specific domains, languages, and language uses.

Due to the way in which training was achieved, most early language models (and still the majority) focus on word-level embeddings. However, embeddings models can also benefit from char-level or character n-gram information to provide a lower-level source of information. This is quite intuitive, because in many cases two words that have same root are likely to be related in some way (for example, beauty and beautiful). This is especially true for some languages like German, where a word can be a concatenation of two others, giving it new meaning (for example, 'Wahrscheinlichkeits' is probability and 'theorie' is theory, thus, 'Wahrscheinlichkeitstheorie' is 'probability theory').

Character n-grams can be very beneficial in a range of different contexts. In the case of morphological rich languages the character n-gram gives an opportunity to assign meaning to novel constructions which may not have been directly encountered in training data. Even if less morphological languages, the character n-gram approach can give information in contexts where novel words are seen perhaps due to the use of foreign words, in the case of neologisms or new word constructions, or even for language errors in text such as where 'baeutifl' might be used in place of the correct word 'beautiful'. Thus, linguistic flexibility and error tolerance are strong advantages of a character n-gram approach, making them natural language units or useful additions to word-level embedding.

Despite this, the relative role and usefulness of character n-grams alongside word level embeddings is still under-studied in the literature – particularly relative to specific tasks. To this end, in this paper we look at the issue of embeddings and n-gram use in the context of an Authorship Attribution task. Here it has been shown previously that the character n-gram model is beneficial to Authorship Attribution (AA) [3–6], but yet the exact reasons behind this success are not clear. While we can assume that low-level features may determine a specific style of writing more precisely than words alone, it is likely that not all character n-grams are equal in importance. This in turn being due to the fact that some n-grams are more frequent than others and give more insights to text meaning. Having this level of knowledge may provide insights into the best design of n-gram level additions to the word level embeddings for particular language variants and task types.

Given the above, in this paper we present a systematic evaluation of the impact of specific character n-grams when applied within a modern transformer encoder model [2]. We do this through the use of a special identifier that indicated char n-gram (so that we can for

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

SAC '23, March 27-March 31, 2023, Tallinn, Estonia

ACM ISBN 978-1-4503-9517-5/23/03.
https://doi.org/10.1145/3555776.3577856

example differentiate ‘act’ in actor from ‘act’ in ‘pact’) for character n-gram tokenizer. We perform this analysis in the context of the news Author Attribution task, and specifically make use of the Russian Taiga KP dataset, and the Reuters_50_50 dataset in English. Our analysis focuses on the issue of character n-gram importance through the use of attention scores.

The contributions of this work are a number of insights into the impact of n-grams on the classification task when used within a complete domain tasks. In particular, the study shows that the character n-grams attention score is higher for n-grams that are considered to be important for authorship identification for humans, and that this is true both for the case of English and Russian datasets. This we suggest underscores the importance of character n-grams as a unit of embeddings in contrast to word only level embeddings, and also that this approach may enhance the usefulness of attention and n-grams in the delivery of explainability or at least interpretability in embeddings models and downstream tasks.

2 RELATED WORK

There have been a number of works that have successfully used either character-based embeddings directly or combination of these with word-level embeddings [7][8][9]. They used RNN, LSTM, and their combinations.

Many works have taken a knowledge driven approach to dataset construction and in particular the tokenizations to be applied. Subword tokenization algorithms rely on the principle that frequently used words should not be split into smaller subwords, while rare words should be decomposed into meaningful subwords. Subword tokenization allows the model to have a reasonable vocabulary size while being able to learn meaningful context-independent representations. In addition, subword tokenization enables the model to process words it has never seen before, by decomposing them into known subwords [10]. There are many subword tokenization techniques which include Byte-Pair Encoding (BPE) [11], WordPiece [12], Unigram [13], SentencePiece [14].

Some papers used character n-gram as a unit of embedding [15][16][17]. There have been a number of works that compare the character n-gram use alongside more traditional encodings. The paper [18] compares the performance of character n-gram features (n = 3–8) and lexical features (unigrams and bigrams of words), as well as their combinations, on the tasks of authorship attribution, author profiling, and discriminating between similar languages.

Turning specifically to the case of author attribution, many papers have looked at the topic of author attribution both at a holistic level, and also from a low-level character n-gram perspective. In [28] the authors examined the author attribution problem in the context of medical topics. Here they manually evaluated which features could reveal a particular author. The range of features identified included: spelling variance, use of capitalisation and abbreviations, the rounding of numbers, use of particular topic words and even particular slang and conjunctive phrases amongst others.

At a machine learning driven level, character n-grams have been identified as the most successful feature in both single domain and cross-domain authorship attribution tasks and various languages, including Russian and Portuguese [4][5][6].

3 METHODOLOGY

In order to provide more insights into the impact of particular character n-grams on the attribution task within an embedding framework, we designed an experiment around the use of attention scores within a transformer context. The use of attention scores as a global measurement of significance of word tokens has been used in a number of works [19-21].

3.1 Datasets

In this work two datasets are used. The first one is `Reuters_50_50` dataset, which contains news articles in English, while the second is the Taiga dataset which contains news articles in Russian. The `The Reuters_50_50` dataset is an archive of over 800,000 manually categorized newswire stories [22]. Taiga [23] is a Russian text corpus where text sources and their meta-information were collected for a number of machine learning tasks. For our purposes a subset of Taiga dataset called Komsomolskaya Pravda (KP) from `taiga_news_all` is used.

To best preserve an author’s style, it was decided to omit extensive preprocessing of the datasets. The only preprocessing that was performed was lowercasing – thus resulting in two variants for each dataset: original text and lowercased text.

3.2 Models

For our character n-grams model we made use of single architecture for both Russian and English language datasets. The model was based on a transformer encoder with scaled dot product attention and positional encoding [27].

Our approach depends on selecting an appropriate approach to content tokenisation. For this purpose we made use of the approach to character n-gram tokenisation that was presented in [4] for both our target language datasets, and used the same char n-grams classification.

4 RESULTS AND ANALYSIS

The goal of this work is to consider the impact of particular character n-gram classes on attentive encoding models. Based on the analysis just presented we proceed by examining the impact of 3-, 4-, and 5-grams on the prediction task. For our analysis of impact, only the top twenty n-grams with both most and least (mostly with negative sign) attention were considered. We believe this is a fair balance which provides insights without overly complicating the analysis process.

Considering first the Reuters dataset, there were many abbreviations among highly-scored n-grams (for 3, 4, and 5 n-grams). Some of them are related to a specific field of interest of an author (e.g., ’HHC’ (Hilton Hotels Corp)) while some of them are more commonplace (e.g., ‘EU ’ (European union)). Analysis also showed that there are some contractions with high attention scores, e.g., "I’ve", "you’r", "’m ve". As for multi term words, the gram ‘-big’ was found (indeed, some Reuters authors write “second biggest” and others write “second-biggest”). Some quite rare words are also encountered among `whole_word` type n-grams, namely, ‘whom’, ‘midst’, ‘bumpy’.

Outside of word and word constituents we can see the influence of style and localisation issues on language use. For example in terms of numbers and money, the use of number deliminator (comma vs decimal point) was a strong feature, as was the use of n-grams related to dates, e.g., ‘/Ju’ (used in “June/July” by some authors), and symbols for measurement ‘-kg’ , ‘-kg_’, ‘_kg’, ‘_cm’, ‘_cm’, and specific scientific symbols. ' C$4', 'a$2', etc.

Turning to the case of Taiga, there was also a lot of abbreviations among highly-scored n-grams (for 3, 4, and 5 n-grams). Again, some of them are related to a specific field of interest of an author and some of them are more common. Interestingly, there were many English (part of) words in the Russian news dataset (in fact, almost all related to words, which are whole_word, multi_word, mid_word, n-grams contains only Latin letters): ’BMW’, ’KIA’, ’USA’, ’CNN’, ’VIP’, ’BBC’, ’CBS’, ’NBC’ (this n-gram is also in English news dataset), ’WADA’, ’CNBC’, and ’NASA’. As for numbers, there was not only Arabic numerals, but also many Roman numbers: ’XIX’, ’III’, ’XVI’, ’XVIII’, ’XIII’, and ’VIII’. Many percentage signs can also be found (inside n-grams): '2-3%', '3,7%', ',65%', '-10%', '. 12%', ' 7,3%', and ' %'.

And some notation can also reveal an author. For example, number notation: ‘№__’ , '№7 ', '№1,', '№5,', or using hashtag '#__' (it was used as "number" word in ‘#89’ or used to denote a hashtag). Some numbers was used with multiple characters, among them: '.51', ':43', ':46', ':43', '/59'. There are some top n-grams related to unit of measurement (of temperature): '3º ', '8° ', '3 °,', '° ', 'C ', '° ', '°-', '3 °,', '-27°' (it was used as opposed to word notation “градусы” – Russian “degrees” that are also found in the text).

As for contractions, some of them are also seen among n-grams. For example, ' г-' ('г-н' is a contraction of word 'господин', the same as 'Mr.' for 'Mister'). Or another example, 'в/ч' (may refer to radiation measurement or a ‘военная часть’ - ‘military unit’). Some letters like: ’š ’, ’ ž’ from other languages were also found among top n-grams.

5 CONCLUSION

By examining issues across both Russian and English language variants, we are confident that this approach has some usefulness across languages and is not as tied to particular morphological features of the language. As such the current work can be seen as work in progress towards that goal. We also note the importance of these features in the general trend towards stylistic tuning in language.

In the context of complex neural architectures based on attention and transformer variants, attention scores give us a simplistic and intuitive means of assessing the importance of individual features. Our analysis showed the importance of a number of non-morphological features such as common abbreviations, contractions, and number use across the target languages. In future work we intend to leverage this information in the construction of hybrid classifiers for models which require textual specificity such as author attribution, as well as to make some changes to our approach to n-gram isolation such that these can be based on the needs of a particular target tasks.

REFERENCES

[1] Jeremy Howard, S. R. (2018). Universal Language Model Fine-tuning for Text Classification. arXiv:1801.06146.

[2] Ashish Vaswani, N. S. (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[3] John Houvardas, E. S. (2006). N-Gram Feature Selection for Authorship Identification. Lecture Notes in Computer Science 4183, 77-86.

[4] Upendra Sapkota, S. B.-y.-G. (2015). Not All Character N-grams Are Created Equal: A Study in Authorship Attribution. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , 93–102.

[5] Tatiana Litvinova, O. L. (2019). Authorship Attribution of Russian Forum Posts with Different Types of N-gram Features. Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval (NLPIR 2019), 9–14.

[6] Ilia Markov1, J. B.-L. (2017). Authorship Attribution in Portuguese Using Character N-grams. Acta Polytechnica Hungarica Vol. 14, No. 3, 59-78.

[7] Yoon Kim, Y. J. (2016). Character-Aware Neural Language Models. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16), 2741-2749.

[8] Lyan Verwimp, J. P. (2017). Character-Word LSTM Language Models. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 417-427.

[9] Piotr Bojanowski, E. G. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, Volume 5, 135–146.

[10] Huggingface. (30 09 2022 .). Summary of the tokenizers. Huggingface: https://huggingface.co/transformers/v4.3.0/tokenizer_summary.html

[11] Rico Sennrich, B. H. (2016). Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1715-1725.

[12] Mike Schuster, K. N. (2012). Japanese and Korean voice search. International Conference on Acoustics, Speech and Signal Processing, IEEE (2012), 5149-5152.

[13] Kudo, T. (2018). Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 66-75.

[14] Taku Kudo, J. R. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71.

[15] John Wieting, M. B. (2016). Charagram: Embedding Words and Sentences via Character n-grams. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1504–1515.

[16] Sho Takase, J. S. (2019). Character n-gram Embeddings to Improve RNN Language Models. The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), 5074-5082.

[17] Tao Shen, T. Z. (2019 03 2019 .). Tensorized Self-Attention: Efficiently Modeling Pairwise and Global Dependencies Together. arXiv:1805.00912v4. Retrieved from arxiv: https://arxiv.org/pdf/1805.00912.pdf

[18] Miguel A. Sanchez-Perez, I. M.-A. (2017). Comparison of Character n-grams and Lexical Features on Author, Gender, and Language Variety Identification on the Same Spanish News Corpus (preprint version). Conference: Experimental IR Meets Multilinguality, Multimodality, and Interaction – 8th International Conference of the CLEF Association (CLEF 2017). Volume: 10456, 145–151.

[19] Xiaobing Sun, W. L. (2020). Understanding Attention for Text Classification. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 3418-3428.

[20] Richard Socher, A. P. (2013). Recursive Deep Models for Semantic Compositionality. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1631-1642.

[21] Andrew L. Maas, R. E. (2011). Learning Word Vectors for Sentiment Analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 142-150.

[22] David D. Lewis, Y. Y. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. Journal of Machine Learning Research 5, 361-397.

[23] Tatiana Shavrina, O. S. (2017). To the methodology of corpus construction for machine learning:“Taiga” syntax tree corpus and parser . PROCEEDINGS OF THE INTERNATIONAL CONFERENCE «CORPUS LINGUISTICS–2017», 78-84.

[24] Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.

[25] He Pengcheng, L. X. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. CoRR abs/2006.03654.

[26] Yuri Kuratov, M. A. (2019). Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019", 1-7.

[27] Canziani, A. (30 09 2022 .). Attention and the Transformer. Retrieved from atcold: https://atcold.github.io/pytorch-Deep-Learning/en/week12/12-3/

[28] Rexha A, K. M. (2018). Authorship identification of documents with high content similarity. Scientometrics, 223-237.