Large-Scale Author Verification: Temporal and Topical Influences

Michiel van Dam
Delft University of Technology
Delft, The Netherlands
M.C.vanDam@student.tudelft.nl

Claudia Hauff
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

ABSTRACT

The task of author verification is concerned with the question whether or not someone is the author of a given piece of text. Algorithms that extract writing style features from texts are used to determine how close in style different documents are. Currently, evaluations of author verification algorithms are restricted to small-scale corpora with usually less than one hundred test cases. In this work, we present a methodology to derive a large-scale author verification corpus based on *Wikipedia Talkpages*. We create a corpus based on English Wikipedia which is significantly larger than existing corpora. We investigate two dimensions on this corpus which so far have not received sufficient attention: the influence of *topic* and the influence of *time* on author verification accuracy.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
General Terms: Experimentation
Keywords: authorship verification, plagiarism detection

1. INTRODUCTION

Plagiarism detection is a field of research whose importance in educational and academic contexts is well-known [5, 12]. Moreover, plagiarism detection has applications in historical and forensic research [3, 15]. Here, it is often necessary to determine the true author of encountered documents. The plagiarism sub-field of *author verification*, which we focus on in this work, frames the question of authorship as a classification task: *Given a piece of text $\mathcal{T}$, an author $\mathcal{A}$ and one or more reference documents known to have been written by $\mathcal{A}$, determine whether $\mathcal{A}$ has written $\mathcal{T}$.*

Authors unconsciously leave clues in their written text, that can identify them as the author. These clues are called *style markers* [4]. There are many different possibilities for style markers, e.g. the average number of words in a sentence, the frequency of words ending in `ing` (for English) and the fraction of times a conjunction is followed by a clarifying statement.

To accurately capture an author's writing style it is important to determine which style markers truly capture the style and which are influenced by the topic of the text or the context. Character N-grams are the most common approach to capture style markers. Not only are they language independent and inexpensive to compute and store, past research has also shown them to be a powerful tool to recognize and verify authors.

A common benchmark for the author verification task is provided by the PAN benchmark series, which started in 2009 and is still ongoing [16, 8]. In 2013 [9], for the first time, multiple languages were evaluated (English, Greek and Spanish) and for each language up to 30 verification problems were provided. For each verification problem, the document to verify, and the author with up to eight short reference documents was provided. *Decreasing* the amount and size of reference documents has been a trend across the years - while initially, whole books were used as reference documents, currently short extracts of books are used. The documents within one verification problem are matched for genre, topic and date of writing.

While this benchmark allows us to compare several algorithms, it is at the same time exceedingly small. This makes it impossible to investigate to what extent currently used style markers are influenced by additional dimensions such as topic and time. In this paper, we address this shortcoming and propose a methodology based on *Wikipedia Talkpages* (WTP), to *automatically* gather a large-scale corpus for author verification, which is controlled for topical and temporal effects.

Having created such a corpus from the English Wikipedia[1], we investigate the following research hypotheses:

Topic hypothesis When using short and diverse sets of reference documents, spanning multiple topics, we hypothesize that the authorship verification problem becomes more difficult and the author verification accuracy drops in comparison to single-topic long reference documents (e.g. books).
An often made assumption in writing style research is that two texts about similar topics will be biased towards being recognized as from the same author, regardless of whether it is correct. We investigate this assumption on our corpus for which the similarities in topic are known.

[1]The corpus is available at `http://www.st.ewi.tudelft.nl/~hauff/wikiAuthorVerificationData.html`.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609504.

Temporal hypothesis It is unrealistic to assume that an author will maintain her writing style across a long period of time. We hypothesize that two texts from the same author will align more in writing style when they have been written within a short time span, compared to two texts written at very different times.

Our work makes the following contributions: (i) we introduce a corpus creation pipeline based on WTPs which can generate corpora with thousands of test cases in many different languages; (ii) we challenge the topic assumption and show that topic alignment improves the accuracy both for the positive and negative class, and (iii) we experimentally show that the writing style of authors changes over time, making two documents from the same author, written a long time apart, more prone to a false classification.

After a short overview of related work, we introduce our corpus creation methodology, then briefly describe the character N-gram based author verification approach we implemented. Finally we report the results when investigating the influences of topic and time on the verification accuracy.

2. RELATED WORK

Different categories of stylistic markers (or stylometric features) have been identified in the past as useful to determine the authorship of a document: lexical (e.g. sentence length, number of unique terms), character (e.g. character N-grams or character types), syntactic (e.g. part-of-speech features), semantic (semantic relationship between words or phrases, e.g. clarification or conjunction) or application specific (e.g. greetings in e-mail messages) [20].

A lot of research has focused on the applicability of different style markers. In [6], 39 text features were compared. It was concluded that word and punctuation mark profiles are strong discriminators between authors, next to character 2- and 3-grams. Stamatatos [20] provides a detailed overview of character attribution methods currently employed, and discusses the usability of different stylistic marker categories. He concludes that character N-grams belong to the most effective features out of all lexical and character style features, while having only minimal computational requirements.

Within the character N-grams technique the most stylistically important features are those N-grams that occur most frequently in a text, indicating an author's preference for specific word endings or conjugations. Stamatatos [19] states that character N-grams *are able to capture complicated stylistic information on the lexical, syntactic, or structural level* and provides examples of lexical ("the", "_to", "tha"), word-class ("ing", "ed_"), and punctuation usage ("._T", "_'T") information in character 3-grams. In [18], character 4-grams were found to be most effective for author attribution.

Ruseti et al. [17] use the one hundred most common character trigrams to distinguish authors, next to a limited set of additional lexical features, achieving more than 95% accuracy on the plagiarism and author unmasking task at PAN 2011. Both works, [17, 19], assume that the most frequent N-grams will consist of stylistic markers, rather than N-grams specific to a topic as writing style is assumed to be topic-independent.

Employing machine learning frameworks to automatically select the best features has been proposed, among others in [1] and [11].

Finally recent author verification attempts are also summarized in the overviews of the various PAN tasks [16], [8], and [9]. For the author verification task it was found that word-based features are useful in distinguishing between authors writing about different subjects. Moreover, most high-scoring submissions used character n-gram features, often combined with other style features.

As stated earlier, the currently used benchmark corpora are rather small with respect to the number of test cases, making investigations into particular dimensions of the problem (such as time or topic) infeasible. At the same time, older corpora contain very long reference documents (entire books), which are not available in many use cases. Recently, Mikros and Perifanos [13] developed an authorship attribution corpus based on tweets. We consider a corpus based on tweets not very suitable for most use cases of author verification as the writing style on Twitter is heavily biased due to the very strong limitations on allowed text. Finally, we also note the use of the Reuters RCV1&2 corpora (newswire texts) for authorship attribution experiments [7]. These corpora however are not controlled for multi-author influences or quotations (as we do). Furthermore, the writing styles of news-wire stories are usually geared towards a particular one as required by the news organization. In comparison, we consider our corpus more in line with use cases that deal with digital forensics.

3. CORPUS CREATION

As previously stated, the trend in authorship research is to move towards evaluating less and shorter texts rather than entire books, and to evaluate algorithms across multiple languages. Therefore we aimed to develop a pipeline that allows us to *automatically* derive large-scale corpora in different languages, with many contributing authors across a range of topics and contributions by authors across an extended period of time.

The input to our pipeline are WTPs. On WTPs Wikipedia contributors leave comments arguing about Wikipedia page changes. WTP comments are annotated with the contributor's user name or IP address and the time of commenting. We chose WTPs over the authors' Wikipedia page contributions, as we expect authors to adhere to the Wikipedia writing guidelines when contributing to an existing Wikipedia page. In contrast, we assume that on WTPs, which do not prescribe to a specific writing style, contributors are more likely to write in their own style.

While in this work we focus on the English portion of Wikipedia, it is evident that such corpora can be built for all languages with sufficient WTP contributions.

WTP Comments.

We consider each Wikipedia contributor as a potential author in our corpus and gather all WTP comments by the contributor. To retrieve the contribution of an author we compare the difference in a WTP between two revisions. Although a revision is created as soon as the WTP is changed (text is added or deleted or reformatted), we only consider those revisions where text is added by a contributor. This approach can lead to some noise (e.g. a contributor may be known under multiple user names or IP addresses, a contribution can contain a quote from another author. etc.), however, a manual analysis confirmed the quality of the derived data. Two hundred contributions between 2,500 and

10,000 characters were examined by the authors: 85% of the contributions had a single author, 5% contained a quotation of a size less than a third of the contribution, 7% contained a larger quotation, and the last 3% contained content otherwise not belonging to the contributing author. Thus, we conclude that our pipeline is of sufficient quality to generate a large-scale corpus for author verification.

For this work we generated a corpus based on a portion (40%) of English Wikipedia[2]. Overall, we observed 2,891 authors with at least four different comments, each one between 2,500 and 10,000 characters long. We chose this length interval in order to create a dataset which, while much larger and more diverse, is similar in spirit to the latest PAN datasets, where reference documents with approximately 1,000 words (∼5,000 characters) are used.

Building Test Sets.

Given the nearly 3,000 (for us) valid Wikipedia contributors, it is evident that we can generate a large number of verification problems (or test cases). In this section, we focus on building a balanced test set to evaluate our topical and temporal hypotheses.

Recall that a test case consists of an author (a Wikipedia contributor $\mathcal{A}$), reference documents (we use **a single** comment by $\mathcal{A}$ as reference document) and an unknown piece of text (positive test case: a comment by $\mathcal{A}$, negative test case: a comment by another contributor).

To investigate our topical hypothesis, we require test cases where the reference and to-be verified documents are similar and dissimilar respectively. To investigate the temporal hypothesis, we require test cases where the reference and to-be verified documents have been written within a short and long period of time respectively.

We generate five test sets. In the following list, *similar* either means similar in topic or similar in time:

- `All annotated`: baseline, containing all test cases
- `Similar`: all test cases where the reference comment is similar to the comment to be verified
- `Different`: all test cases where the reference comment is dissimilar to the comment to be verified
- `Matching`: all test cases where similarity occurs when the reference and to-be-verified comments are by the same author
- `Nonmatching`: all test cases where similarity occurs when the reference and to-be-verified comments are from different authors

To generate test cases with known topical similarity, we determined the semantic similarity between the two Wikipedia article pages that the WTPs from which a pair of comments were derived, are attached to. The similarity score (between 0 and 1) was computed with the Wikipedia Miner Toolkit [14]. Comment pairs with a score above 0 were considered similar, comment pairs with a score of 0 as dissimilar.

To generate test cases across different time spans, we simply made use of the timestamps attached to each comment. Comments written more than three years apart are considered dissimilar in time. As similar in time we consider comments that have been written less than a week apart in time.

[2]We also experimented with the French and German Wikipedia, which yielded similar results. Due to space constraints, we focus on English Wikipedia in this work.

Having set the similarity thresholds, we now derive the test cases. First, those authors are selected that have both similar and dissimilar WTP comments. For each author a test case is created with two similar posts, and a test case with dissimilar posts.

After these same-author test cases have been generated for each author, for every author two different-author test cases are generated: one for similar time/topic, and one for dissimilar time/topic. The text from the unknown author is selected by taking a random text from the available comments, and verifying that it conforms to the time/topic requirements.

This method yields four test cases for every author: one same-author similar time/topic, one same-author dissimilar time/topic, one different-author similar time/topic, and one different-author dissimilar time/topic. The precise number of generated test cases is presented in Section 4.

Finally, each generated test set is balanced to having a 50% same-author rate by randomly discarding test cases if necessary.

3.1 Character N-gram approach

We now briefly describe the character N-gram approach we implemented for our experiments. On a high level, the algorithms works as follows: first, all documents are preprocessed. Then, author profiles are created for the set of known reference documents and for the unknown document, based on the Common N-Grams (CNG) approach[10]. Essentially, each profile is a vector of N-grams. The distance between the known author profile and the unknown author profile is calculated according to [21], and finally based on the computed distances a judgement is made whether the unknown author is the same person as the known author.

The pre-processing steps included simple text transformations such as replacing all digits by a special symbol, since the important stylistic information is the use of digits rather than the exact combination of digits as shown in [7]. Previous work [19] concluded that for the CNG method N-grams of length three to five and a profile length of between 1,000 and 5,000 N-grams usually gives the best results. Our preliminary experiments confirmed those results, and we fixed $N = 4$ for all reported experiments.

The presented algorithm calculates how distinct authors are from each other, using a distance measure for the character N-gram profiles. The average distance of *all* test cases is taken as threshold for similarity. This forces roughly half of the test cases to be classified as positive (same author) and half as negative (different author).

4. EXPERIMENTS

Topical Similarity.

In total, there are 2359 test cases with a reported topic similarity of 0, and 1924 test cases with a reported similarity > 0. The test case numbers in Table 1 reflect this skew; since all test sets are balanced there are far fewer test cases in the `Similar` test set than the others. Despite this, even the smallest generated test set is still ten times larger than last year's PAN dataset.

In Table 1 we also report the number of test cases generated for each of the five test sets as well as the accuracy achieved by the CNG algorithm. If our topic hypothesis would hold, then cases where the topic is the same should

Test set	# Test cases	Accuracy
All annotated	3950	0.598
Similar	316	0.642
Different	2406	0.580
Matching	1544	0.618
Nonmatching	1240	0.588

Table 1: Number of test cases and the accuracy of the CNG algorithm when comparing comments about similar topics to comments about dissimilar topics. A random baseline achieves accuracy 0.5.

Test set	# Test cases	Accuracy
All annotated	1368	0,633
Similar (<1wk)	684	0,665
Different (>3yr)	684	0,588
Matching	684	0,684
Nonmatching	684	0,580

Table 2: The recorded accuracy when comparing comments made within one week to comments made more than three years apart.

be ranked as being the same author more often, than cases where the topic is different. As this does not influence the `Similar` and `Different` test sets, we would expect the reported accuracy not to diverge too much from the baseline accuracy (`All annotated`). In contrast, for the `Matching` test set we expect a higher reported accuracy than for the baseline, while for the `Nonmatching` test set we expect a lower reported accuracy than for the baseline.

We observe that the second expectation holds, however, the first expectation is not reflected in the results. The `Similar` test set yields a higher accuracy than the baseline, while the `Different` test set yields a lower accuracy. We thus have to conclude that topic words are indeed also significant style markers that play a role in the author verification task.

Temporal Similarity.

The results of the test sets generated according to temporal similarity are shown in Table 2. The first observation to make is that the `Similar` test set performs better than the baseline (`All annotated`) test set. The reported accuracies suggest that the writing style changes over time - the accuracy of the `Similar` test set (0.67) is considerably higher than the accuracy of the `Different` test set (0.58). A similar observation can be made when comparing `Matching` and `Nonmatching`. The conclusion we draw from these experiments is that time matters. The writing style changes over time, making two texts written far apart in time more difficult to link to the same author.

In [2] it was already suggested that writing style changes over time. This is confirmed by the our results. However, in [2] it is also suggested that a higher time gap could be a reason for a better discrimination between old and new works. We performed additional experiments and generated an additional test set with a one year time gap. However, the results did not confirm this claim - the results for the one and three year time gap were roughly the same. Due to data sparsity, time gaps larger than three years could not be investigated.

5. CONCLUSIONS

In this work, we proposed a methodology that allows us to generate large-scale corpora for author verification across multiple languages, based on Wikipedia Talkpages. We introduced two ways of building test cases to investigate topical and temporal hypotheses and reported results for the English portion of Wikipedia.

We investigated the topical influence and found that the topic does indeed influence the verification accuracy. The influence is not simply that similar topics skew the results towards false positives. Rather, similar topics overall (positive and negative) cases were found to increase accuracy of author verification.

We also investigated the influence of time and found that writing style indeed changes over time, by comparing WTP contributions made within a week with WTP contributions made years apart. Author verification is more accurate when comparing texts that have been written within a short period of time.

In the future, we plan to investigate more languages - with the corpus generation pipeline in place, large-scale corpora for all languages with sufficient WTPs can be created. This will allow us to investigate the influence of the language family on the author verification accuracy. Furthermore, while in this work we restricted our investigation to the CNG algorithm (which is language independent), we will also consider additional language-dependent algorithms.

6. REFERENCES

[1] S. Argamon, M. Šarić, and S. Stein. Style mining of electronic messages for multiple authorship discrimination: first results. In *SIGKDD '03.*, pages 475–480, 2003.

[2] F. Can and J. M. Patton. Change of Writing Style with Time. *Computers and the Humanities*, 38(1):61–82, 2004.

[3] C. E. Chaski. Who's at the keyboard? authorship attribution in digital evidence investigations. *IJDE*, 4(1), 2005.

[4] N. Cheng, R. Chandramouli, and K. Subbalakshmi. Author gender identification from text. *Digital Investigation*, 8(1):78–88, 2011.

[5] R. Clarke. Plagiarism by academics: More complex than it seems. *Journal of the Association for Information Systems*, 7:91–121, 2006.

[6] J. Grieve. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3):251–270, 2007.

[7] J. Houvardas and E. Stamatatos. N-gram feature selection for authorship identification. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, pages 77–86, 2006.

[8] P. Juola. An Overview of the Traditional Authorship Attribution Subtask. *CLEF 2012*, 2012.

[9] P. Juola and E. Stamatatos. Overview of the Author Identification Task at PAN 2013. *CLEF 2013*, 2013.

[10] V. Kešelj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 255–264, 2003.

[11] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. *ICML '04*, 2004.

[12] H. A. Maurer, F. Kappe, and B. Zaka. Plagiarism-A Survey. *Journal of Universal Computer Science*, 12(8):1050–1084, 2006.

[13] G. K. Mikros and K. Perifanos. Authorship attribution in greek tweets using author's multilevel n-gram profiles. In *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[14] D. Milne and I. H. Witten. An open-source toolkit for mining Wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[15] F. Mosteller and D. L. Wallace. Inference in an Authorship Problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[16] M. Potthast, A. Eiselt, A. Barrón-Cedeño, B. Stein, and P. Rosso. Overview of the 3rd International Competition on Plagiarism Detection. *CLEF 2011*, 2011.

[17] S. Ruseti and T. Rebedea. Authorship Identification Using a Reduced Set of Linguistic Features. *Notebook for PAN at CLEF 2012*, 2012.

[18] C. Sanderson and S. Guenter. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In *EMNLP '06*, pages 482–491, 2006.

[19] E. Stamatatos. Author Identification Using Imbalanced and Limited Training Texts. In *DEXA '07*, pages 237–241, 2007.

[20] E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[21] E. Stamatatos. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In *3rd PAN Workshop*, pages 38–46, 2009.