Author Masking through Translation

Notebook for PAN at CLEF 2016

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder

Dhirubhai Ambani Institute of Information and Communication Technology
{yashwant.keswani, harshjtrivedi94, parth.mehta126, prasenjit.majumder}@gmail.com

Abstract This notebook paper documents the approach adopted by our team for Author Masking Task in PAN 2016. For the purpose of masking the identity of the author, we use a simple translation based approach. From the source language (English), the text is translated to an intermediate language before it gets finally translated back to English. In this process, depending on the translation model and various penalties used during the translation process, a change of the structure of the language seeps in. Besides this, translation process can also change the vocabulary used in the text as well as the average sentence length. We attempt to use this approach for obfuscating the identity of author of the text.

1 Introduction & Related Work

Tthe task is as follows [1]: Given a document, we need to reformulate the text so as to satisfy the following conditions:

1. Mask the identity of the original author of the document so that state of art authorship attribution systems can not identify the author.
2. Retain maximum content of the document so the meaning of the text does not change
3. Maintain the linguistic quality of the document so that a human cannot identify it to be machine generated text.

Training set consists of sets of (3-5) documents written by n authors: $A_1, A_2...A_n$. The test set consists of one or more documents, corresponding to each of these authors, which needs to be masked: $D_{i_{original}}$ for author A_i. The documents $D_{i_1}, D_{i_3}...D_{i_m}$ represent the writing style of A_i. Here, m $\in \{3, 4, 5\}$ is the number of documents written by author A_i provided in training set. Using this information about writing style of A_i, the task is to rewrite / paraphrase the document $D_{i_{original}}$ to $D_{i_{obfuscated}}$ in such a way, that evaluator algorithms do not identify A_i as the author of $D_{i_{obfuscated}}$ or equivalently, these algorithms should not detect that author of $D_{i_{obfuscated}}$ and $D_{i_1}, D_{i_3}...D_{i_m}$ is same.

The task of authorship masking has not been well researched. Patrick Juola & Darren Vescovi used Brennan-Greenstadt Obfuscation corpus with JGAAP systems to test different methods of authorship attribution against text written in deliberate attempt to obfuscate the style [2]. Gary Kacmarcik & Michael Gamon have explored techniques for reducing the effectiveness of standard authorship attribution techniques so that an

author can preserve the anonymity for a particular document [3]. They have discussed method of identifying most salient features for identification and shown how this information can be fed back to create the obfuscated document so that the attribution moves away from the original. Also, there has been a previous attempt to perform this task by to-fro language translation: English → French → English. As mentioned in their paper [4], considering the low quality of the state-of-the-art translation methods then, they were not able to yield a good performance. In this attempt, we try to test the idea of to-fro translation using an additional intermediate language and check its performance with current state-of-the-art translation tools.

2 Approach

Our approach tries to exploit the corruption caused by translation system when translating a piece of text from one language to another and leverage this to perform the task of obfuscation. The idea is to perform sequential translation of the to-be-obfuscated document of each author to a few intermediate languages and then translate it back to English: English → IL_1 → IL_2 → ... IL_n → English, where IL_j is the intermediate languages. Our initial approach was to use the translation API provided by Google Translate[1]. However Google translate uses English as a pivot language for translation. Which means while translating a document from English to French to German, the English document will be translated to French, which will be translated back to English, and the new English document will be then translated to German. This approach didn't turn out to be much useful. Most machine translation systems don't drift to a new sentence while translating between two pairs of languages. Which means translating a English sentence to French and then back to English will, in most cases, return the original English sentence itself. To counter this we tried using other translation systems like Yandex[2] and Microsoft Bing Translate[3] for performing a part of the intermediate translations. For example we would translate a English sentence to French using Google translate and then use Bing translate to get the German sentence which will then be translated to English using Yandex. This approach seemed promising in terms of language quality. Most of the sentences generated were human readable with few phrase positioning shifts and some words being replaced by the synonyms. However there were certain unexpected errors deploying and running the software on TIRA platform [5] [6] due to a high number of API calls required by such a system. Finally we opted to train our own translation models using Moses SMT toolkit[7]. We used German and French as the intermediate languages: English → German → French → English.

We used the Europarl corpus[8] for training. The Europarl corpus provides parallel translations where one of the languages is English. We created the German-French corpus using the English-German and English-French corpus. We used a randomly selected subset of 100K sentences per language pair for training the translation model and tuned the model with another 5K randomly sampled sentences.

[1] `https://translate.google.com`
[2] `https://translate.yandex.com/`
[3] `https://www.bing.com/translator`

3 Evaluation Details

The organizers have proposed the following parameters for evaluation of Author Masking task. The quality of submitted systems is judged based on:

- Safeness: does forensic analysis reveal the original author of its obfuscated texts?
- Soundness: are the obfuscated texts textually entailed with their originals?
- Sensibleness: are the obfuscated texts inconspicuous to a human reader?

Safeness was evaluated for all the three participant teams by the task organizers [1]. It was done by a large-scale safety evaluation of the three participant obfuscation approaches. This was done based on the authorship verification approaches submitted to the PAN 2013-2015 authorship verification tasks [9,10,11]. These authorship verification approaches were run on obfuscated versions of the test datasets of their respective years. Each submitted obfuscation approach was run, yielding three additional test datasets per original dataset. Only the English portions of the test datasets were used.

The following table documents the average performance drop in terms of final score of the verification approaches when compared to their performance on the unobfuscated (original) test datasets. Higher the performance drop, better the obfuscation. Our team stood second in terms of the safeness measure.

Table 1. Average performance drops in terms of "final scores" of the authorship verifiers submitted at PAN 2013 to PAN 2015 when run on obfuscated versions of the corresponding test datasets as per the submitted obfuscators.

Participant	**PAN 2013**	**PAN 2014 EE**	**PAN 2014 EN**	**PAN 2015**
Mihaylova *et al.* [12]	-0.10	-0.13	-0.16	-0.11
Keswani *et al.*	-0.09	-0.11	-0.12	-0.06
Mansoorizadeh *et al.* [13]	-0.05	-0.04	-0.03	-0.04

The organizer have acknowledged the gap in automatic evaluation measures for this task and have invited proposals for an automatic evaluation measure, which is done through a separate task, "Obfuscation Evaluation". Results of all three "Author Masking" teams in terms of soundness and sensibleness would be available in the task notebooks of "Obfuscation Evaluation" teams.

4 Conclusion & Future Work

Overall the use of machine translation systems seems a worthy attempt at Authorship attribution. We would like to try several further approaches in future. For instance, due to the limitations of the virtual machines, we had to reduce the size of the training corpus. We would like to see the effect of using the entire Europarl Corpus(1.5 million sentences). We would also like to try it on a different corpus which has a more broader vocabulary. Another approach we would like to further explore is tuning the language

model and sentence length penalties in moses translation system. These penalties control the linguistic quality and length of the translated sentences. Yet another possibility is to use the word usage trends to manipulate the translations. Replacing a few words that are used in recent times by those that were popular in 18th century would be an interesting approach.

References

1. Martin Potthast, Matthias Hagen, and Benno Stein. Author Obfuscation: Attacking State-of-the-Art Authorship Verification Approaches. In *Working Notes Papers of the CLEF 2016 Evaluation Labs*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, September 2016.
2. Gary Kacmarcik and Michael Gamon. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451. Association for Computational Linguistics, 2006.
3. Patrick Juola and Darren Vescovi. Empirical evaluation of authorship obfuscation using jgaap. In *Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security*, pages 14–18. ACM, 2010.
4. Josyula R Rao, Pankaj Rohatgi, et al. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*, 2000.
5. Tim Gollub, Benno Stein, Steven Burrows, and Dennis Hoppe. TIRA: Configuring, Executing, and Disseminating Information Retrieval Experiments. In A Min Tjoa, Stephen Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *9th International Workshop on Text-based Information Retrieval (TIR 12) at DEXA*, pages 151–155, Los Alamitos, California, September 2012. IEEE.
6. Martin Potthast, Tim Gollub, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury, and Elaine Toms, editors, *Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14)*, pages 268–299, Berlin Heidelberg New York, September 2014. Springer.
7. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.
8. Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.
9. Patrick Juola and Efstathios Stamatatos. Overview of the author identification task at pan 2013. In *CLEF (Working Notes)*, 2013.
10. Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A Sanchez-Perez, and Alberto Barrón-Cedeño. Overview of the author identification task at pan 2014. In *CLEF (Working Notes)*, pages 877–897, 2014.
11. Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. Overview of the pan/clef 2015 evaluation lab. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 518–538. Springer, 2015.
12. Tsvetomila Mihaylova, Georgi Karadjov, Preslav Nakov, Yasen Kiprov, Georgi Georgiev, and Ivan Koychev. SU@PAN'2016: Author Obfuscation—Notebook for PAN at CLEF 2016. In Balog et al. [14].

13. Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, and Mahdy Eskandari. Author Obfuscation using WordNet and Language Models—Notebook for PAN at CLEF 2016. In Balog et al. [14].
14. Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*, CEUR Workshop Proceedings. CEUR-WS.org, 2016.