Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text

Aylin Caliskan
Department of Computer Science
Drexel University
Philadelphia, USA
Email: ac993@cs.drexel.edu

Rachel Greenstadt
Department of Computer Science
Drexel University
Philadelphia, USA
Email: greenie@cs.drexel.edu

Abstract—**In this paper, we investigate the effects of machine translation tools on translated texts and the accuracy of authorship and translator attribution of translated texts. We show that the more translation performed on a text by a specific machine translation tool, the more effects unique to that translator are observed. We also propose a novel method to perform machine translator and authorship attribution of translated texts using a feature set that led to 91.13% and 91.54% accuracy on average, respectively. We claim that the features leading to highest accuracy in translator attribution are translator-dependent features and that even though translator-effect-heavy features are present in translated text, we can still succeed in authorship attribution. These findings demonstrate that stylometric features of the original text are preserved at some level despite multiple consequent translations and the introduction of translator-dependent features. The main contribution of our work is the discovery of a feature set used to accurately perform both translator and authorship attribution on a corpus of diverse topics from the twenty-first century, which has been consequently translated multiple times using machine translation tools.**

Index Terms—**authorship attribution; privacy; anonymity; machine translation; machine learning**

I. Introduction

Stylometry is a field that relies on linguistic information found in a document to perform authorship recognition. Stylometry is currently used in intelligence analysis and forensics. The 2009 Technology Assessment for the State of the Art Biometrics Excellence Roadmap (SABER) commissioned by the FBI stated that, "As non-handwritten communications become more prevalent, such as blogging, text messaging and emails, there is a growing need to identify writers not by their written script, but by analysis of the typed content [1]." Authorship attribution is the problem of determining a text's author, which we seek to accomplish using stylometric analysis. We show that authorship attribution can still be achieved in translated texts using a set of features, indicating that the authors are not obfuscated. This is a serious privacy concern that prevents anonymous speech.

Rao and Rohatgi [2] introduced the idea of translating text to a different language and then back to its original language using a machine translation tool to obfuscate a text's author. Translated text accumulates properties from the machine translation tool, which is called the translator effect. The translator effect introduces an extra author to the translated text, which is the machine translation tool itself. We train a classifier to detect the machine translation tools' footprints to attribute a translator to the anonymous text. We show that the translator effect does not prevent authorship attribution even though the translator introduces new features to the text.

Machine translators are categorized by the techniques they use to perform translations. Bing's[1] and Google's[2] translators both rely on statistical machine translation. When two translators use the same technique, as is the case with Bing's and Google's translators' statistical machine translation, they do not produce the same output given the same input. Because of these differing translator effects, we can use certain features to enable the attribution of translators as discussed in the experiments section.

II. Related work

State-of-the-art stylometry methods can identify individuals in sets of 50 authors with over 90% accuracy as shown in Abbasi and Chen's work [3]. There has not been much research on identifying the translator effect, translators, and authors in translated text. Suresh et al. [4] were able to match the translated text with the machine translation tool used in the translation of the original text.

Hedegaard and Simonsen [5] researched authorship attribution in translated text, which we outperform in this work. They used classifiers based on frame semantics to discover whether adding semantic features to lexical and syntactic features would improve authorship attribution. Their studies were conducted on a corpus that had a limited number of authors from a specific time period and cultural context, which had only undergone one-way translation.

[1] http://www.microsofttranslator.com/

[2] http://translate.google.com/

III. Corpus selection

Data selection for authorship attribution is an important step. Luyckx and Daelemans [6] show that the number of authors and the amount of text have a big impact on the efficiency of classification. We used the Brennan-Greenstadt Adversarial Stylometry Corpus [7], which has thirteen authors and a minimum of 5000 words per author. All writing samples are written in English by native English speakers. The adversarial stylometry corpus includes one obfuscation and one imitation document per author besides the author's original writing. We excluded these adversarial documents in order to experiment only with the original writing styles. We were left with thirteen authors, 126 documents containing an average of 4933 words per author and 500 words per document. Forsyth and Holmes [8] show that a minimum of 250 words is required in text for authorship attribution. We tested authorship attribution accuracy on a range of datasets with documents varying from 400 to 600 words. 500 words per document led to highest accuracy and accordingly, we divided the datasets into 500 word chunks. The writing samples in the corpus have random topics and therefore are not content-dependent. Schein and Caver [9] show that attribution markers are influenced heavily by topics and effect the authorship attribution rate. We aim to avoid this effect by having varying topics among texts and authors.

In order to create the machine-translated texts, we applied three different sequences of translations to the original corpus by both Bing's and Google's translators. The first sequence translated the original texts to German and then back to English. The second sequence translated the original texts to Japanese and then back to English. The third and last sequence translated the original texts to Japanese, then to German, and then back to English. Hedegaard and Simonsen used eighteen documents in their translator attribution experiments. Their corpus consists of English translations of 19^{th} century Russian romantic literature. The experimented texts have three authors and three translators whereas ours has thirteen authors, two statistical machine translators and two or three consequent translations. Our corpus has more translator effect in the translated text due to two or three consequent translations compared to their single translation from Russian to English. Additionally, our corpus consists of modern text of diverse topics written in the 21^{st} century. As a result, our corpus is more diverse and current.

After performing the two-way translation experiments, we tested the validity of our feature set on one-way translations. We used the work of two French authors and four Dutch authors from the Ad-hoc Authorship Attribution Competition[3] dataset. We translated these texts to English by using Google Translate, Language Weaver and Systran. Both Language Weaver and Systran are also statistical machine translation tools. Each Dutch document ranged from 400 to 600 words. The Dutch dataset consisted of essays on the same six topics by the four authors and is therefore topic-dependent. Because of these qualities in the Dutch dataset, we are able to observe the effect of document length and topic-dependency on authorship and translator attribution.

French and Dutch belong to different language families, namely Romance and Germanic, and therefore possess different grammatical structures. This distinction between the two languages allows us the opportunity to compare language family independent features in translator and authorship attribution.

IV. Experiment design

We divided our experiments into two categories; namely, translator and authorship attribution. We experimented using a variety of features to identify the features leading to highest accuracy as measured by the true-positive rate. We utilized two authorship attribution tools, (1) JGAAP[4] developed by Juola et al. [10] and (2) JStylo, a novel framework for authorship attribution that was developed by McDonald et al.[11]. JGAAP is limited to analysis using one feature at a time. We performed the majority of our experiments using JStylo, which is capable of using a set of multiple features.

A. Translator attribution experiments

To prepare the texts in our corpus, we normalized the documents by stripping all non-ASCII and non-printing characters while preserving the whitespace. Using JGAAP, we trained a Naïve Bayesian classifier and a support vector machine with sequential minimal optimization (SMO) based on Platt's [12] method. We trained the classifiers individually by character grams, part-of-speech (POS) tags, word grams, word lengths, words, function words, sentence length and rare words. We extracted the features with the most frequent and the least frequent events. These features were extracted from documents that were translated to German and then back to English. We tested translator attribution on these documents by using a part of these documents as the training set and the rest as the testing set.

In JStylo, we experimented with several feature combinations and selected those feature sets resulting in high accuracy. The first high accuracy yielding feature set was the '9-Feature Set' used by Brennan and Greenstadt [7].

The extracted '9-Feature Set' was classified with SMO using a polynomial kernel by running 10-fold cross-validation. The experiments were performed on a combination of datasets using the Google (*google*) or Bing (*bing*) translations where the suffixes *en*, *de*, and *ja* correspond to English, German, and Japanese translations, respectively.

The results of these experiments was the attribution of a text as being translated either by *google* or *bing*. We experimented using combinations of features from the '9-Feature Set' and the 'WritePrints Feature Set', which is a partial set of features used by Li et al. [13].

After many possible permutations of feature selection for our set, we selected the set yielding the highest accuracy, as

[3] http://www.mathcs.duq.edu/ juola/authorship_materials2.html

[4] http://evllabs.com/jgaap/w

shown in Table I, which will be called ‘Translation Feature Set’.

Translation Feature Set
Average characters per word
Character count
Function words
Letters
Punctuation
Special characters
Top letter bigrams
Top letter trigrams
Words
Word lengths

TABLE I
TRANSLATION FEATURE SET

For the ‘functions words’ feature, we used the 512 common function words used by Koppel et al. [14]. For feature classes with many features, such as character bigrams, we limited the class to the top 50 extracted features. These features were also classified with SMO using a polynomial kernel by running 10-fold cross-validation resulting with translator classification as *google* or *bing*.

We also tested the ‘Translation Feature Set’ on one-way French and Dutch translations performed with Google Translate, Language Weaver and Systran. The experiments used the exact same methods as the two-way translation translator attribution experiments.

B. Authorship attribution experiments

We followed the same approach in JGAAP as we did in the translator attribution experiments to attribute authors to the anonymous documents that went through one-way and two-way translations.

In JStylo, we tested with all the possible features that were used in the translator attribution experiments with the same machine learning methods by training on the original English writing samples of thirteen authors. The highest accuracy yielding feature set was again the ‘Translation Feature Set’.

V. RESULTS AND EVALUATION

A. Translator attribution results

Our results support Hedegaard and Simonsen’s [5] suggestion of combining features to increase attribution accuracy. Using a single feature at a time had less successful classification accuracy. JStylo experiments using the ‘Translation Feature Set’ had on average 7% better correct classification than experiments using the ‘9-Feature Set’. The results of the JStylo experiments using the ‘Translation Feature Set’ are as shown in Table II.

Dataset	Correct Attribution
en_de	90.87%
en_ja	98.80%
en_ja_de	98.81%
en_de & en_ja	90.44%
en_de & en_ja &en_ja_de	91.13%

TABLE II
‘TRANSLATION FEATURE SET’ TRANSLATOR ATTRIBUTION

Our translator attribution results showed higher accuracy for Japanese translations than for German translations. This indicates that Google’s and Bing’s Japanese translations are less similar than their German translations. Texts which had undergone the most iterations of translations were classified with the highest accuracy, validating our hypothesis that the more consequent translations performed on a text, the stronger the translator footprint will become. Our results showed on average 91.13% accuracy for translator attribution.

B. Translator attribution results in one-way translations

Dataset	Correct Attribution
french_translators	92.75%
dutch_translators	94.44%

TABLE III
‘TRANSLATION FEATURE SET’ TRANSLATOR ATTRIBUTION ON ONE-WAY TRANSLATIONS

‘Translation Feature Set’ led to the highest accuracy rate in attributing Google Translate, Language Weaver and Systran, which are as shown in Table III. All other possible feature sets available in JStylo led to lower accuracy rates than the ‘Translation Feature Set’.

C. Authorship attribution results

Using a single feature at a time resulted in a correct classification rate close to the random chance rate of 7.69%. JStylo experiments using the ‘9-Feature Set’ had on average a 16% less correct classification rate than experiments using the ‘Translation Feature Set’ in Table I, as shown in Table IV. The original writing samples were classified with 97.62% accuracy, labeled as *original_text* in Table IV.

Dataset	Correct Attribution
en_de_bing	96.83%
en_de_google	97.62%
en_ja_bing	100.00%
en_ja_google	89.68%
en_ja_de_bing	77.78%
en_ja_de_google	87.30%
all_bing	91.54%
all_google	91.53%
all_translations	91.54%
original_text	97.62%

TABLE IV
‘TRANSLATION FEATURE SET’ AUTHORSHIP ATTRIBUTION

For authorship attribution in translated text, combining several features also led to higher accuracy than using a single classifier as was the case for translator attribution. Hedegaard and Simonsen argue that “[f]or translated texts, a combined method of frequent words and frames can outperform methods based solely on traditional markers, on translated texts.” Our results show that we outperform Hedegaard and Simonsen’s results using traditional markers in the ‘Translation Feature Set’ shown in Table I without using context-related features such as semantic frames. We are able to achieve this despite an increased translator effect in our corpus which contains documents from consequent translations of different languages. We can identify an author with 91.54% accuracy on average compared to Hedegaard and Simonsen’s average authorship attribution accuracy of 75.27% using their proposed feature set.

Hedegaard and Simonsen also suggest that if semantic markers are not used, authorship attribution may not be possible because of the translator footprint. Our dataset with the most translation iterations was affected thrice by the translator and had the lowest authorship attribution accuracy, demonstrating the validity of Hedegaard and Simonsen’s argument. A broader study on translator attribution and authorship attribution in translated text which includes semantic features may be conducted if the accuracy continues to decrease as the number of consequent translations on a single document increases. The results of such a survey will depend on the translator’s ability to preserve semantics.

Subsequent to our experiments and after discovering the ‘Translation Feature Set’ shown in Table I to yield the highest accuracy for both translator and authorship attribution, we calculated the information gain of the extracted features using WEKA [15]. The comparison of the effectiveness of these features in translator attribution vs. authorship attribution of translated text is as shown in Figure 1.

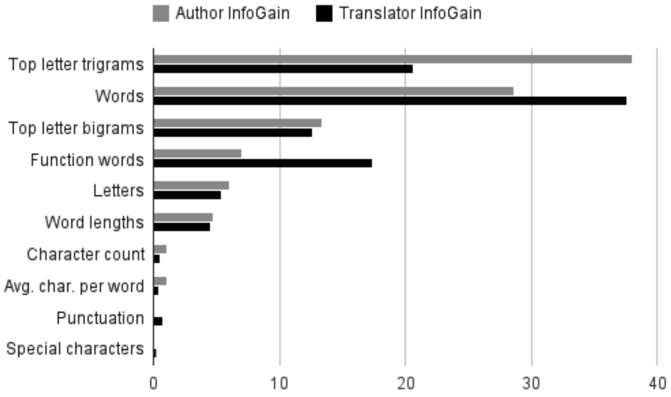

Fig. 1. Comparison of the effectiveness of ‘Translation Feature Set’ in translator attribution vs. authorship attribution

Using the results shown in Figure 1, we can distinguish between more translator-dependent features and more preserved stylometric features in translated text.

The preserved stylometric features in descending effect order are mainly: top letter trigrams, words, top letter bigrams; less effectively: function words, letters, and word lengths. Character-count and characters-per-word had a little effect, while punctuation and special characters had no effect. Translator-dependent features in descending effect order are mainly: words, top letter trigrams, function words, and top letter bigrams; less effectively: letters and word lengths. Character-count, characters-per-word, punctuation, and special characters had little effect.

The comparison shown in Figure 1 demonstrates that ‘function words’ are translator-effect-heavy, but less important for authorship attribution. Hedegaard and Simonsen also argues that the impact of the translator may add sufficient noise to render authorship attribution in translated text very difficult. Consequently, excluding a translator-effect-heavy feature such as ‘function words’ should improve authorship attribution in translated text. To test this claim, we performed an additional experiment in which we excluded function words from our ‘Translation Feature Set’ and compared it to our original authorship attribution results. The results of this experiment are as shown in Figure 2.

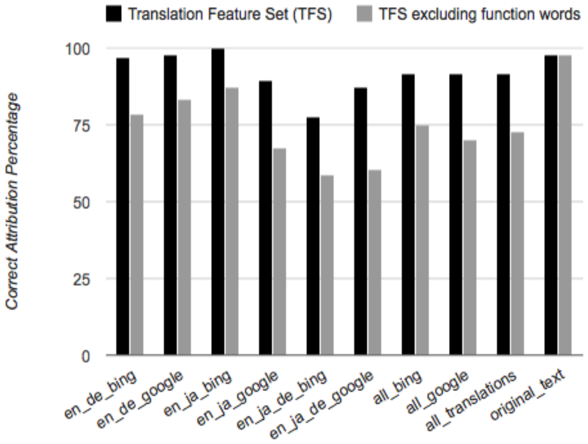

Fig. 2. Comparison of authorship attribution using the ‘Translation Feature Set’ and excluding function words

The results shown in Figure 2 demonstrate that excluding the translator-effect-heavy feature ‘function words’ does not improve authorship attribution. In fact, there is a noticeable decrease in the correct classification rate when ‘function words’ is excluded, suggesting that translator-effect-heavy features do not interfere and can actually aid in more accurate authorship attribution. However, a deeper study regarding the effects of such features is necessary to arrive at a clearer conclusion.

D. Authorship attribution results in one-way translations

Dataset	Correct Attribution
french_google	100.00%
french_languageweaver	100.00%
french_systran	100.00%
dutch_google	60.42%
dutch_languageweaver	70.83%
dutch_systran	75.00%

TABLE V
‘TRANSLATION FEATURE SET’ AUTHORSHIP ATTRIBUTION ON ONE-WAY TRANSLATIONS

'Translation Feature Set' led to the highest accuracy rate in attributing French and Dutch authors using their texts translated to English, which are as shown in Table V. All other possible feature sets available in JStylo led to lower accuracy rates than the 'Translation Feature Set'. Authorship attribution accuracy of Dutch authors is significantly lower than authorship attribution accuracy of all other authors. As described in the 'Corpus selection' section, the Dutch dataset possesses different qualities than the datasets of the other languages. Firstly, the documents have a varying size between 400 and 600 words, whereas the documents of the other datasets are closer to 500 words, which is the optimum length of a document for authorship attribution purposes within JStylo. Additionally, the essays from each author are on the same six topics: the TV show 'Big Brother', smoking, football, a children's story, 'Red Riding Hood', and a historical tale. As a result, word choices and sentences between the authors' essays are very similar. This topic-dependency makes authorship attribution a more difficult task.

Afroz et al. [16] show that authorship attribution is inhibited if an author is trying to imitate another author. Since all the Dutch authors' essays are about existing stories or cases, they may have been influenced by a dominating style. Such an effect may cause some degree of imitation and render authorship attribution difficult. Also, character count is one of the features in the 'Translation Feature Set' and it depends on the length of the document. Since we are using it in the Dutch dataset, it causes a misleading effect because of the varying document sizes in this dataset.

VI. Conclusions and future work

Machine translation tools introduce an effect on translated text that allow for identifying the machine translation tool used to translate the text. In future research, we would expect a higher accuracy in translator attribution among a mixture of rule-based machine translators, statistical machine translators, example-based machine translators, and hybrid machine translators; since they will have more diverse footprints compared to a subset of statistical machine translator footprints.

Authorship attribution of translated text is successful given the existence of a translator effect on the text. Translated texts preserve some of original texts' stylometric features. The more a text goes through iterations of translations, the less preserved the stylometric features become. In future work, we can translate texts for more than three iterations to identify the number of consequent translations necessary to render their authors anonymous. We also want to further investigate authorship attribution in one-way translations of different document lengths. Document length has an impact on attribution accuracy. Formulating the optimum text length for different languages or topics would present a guide for future research.

Machine translation tool attribution and authorship attribution share similar effective features albeit in differing importance levels. These features need to be present in both translator attribution and authorship attribution since they improve attribution accuracy and result in decreased accuracy when removed even though a certain feature may be more effective for either translator or authorship attribution.

Acknowledgment

We thank DARPA (grant N10AP20014) for supporting this work. We are grateful to Sadia Afroz, Seha Islam, and Ariel Stolerman for their invaluable suggestions on the project. We are thankful to Michael Brennan for providing the two-way translations.

References

[1] J. Wayman, N. Orlans, Q. Hu, F. Goodman, A. Ulrich, and V. Valencia, "Technology assessment for the state of the art biometrics excellence roadmap," http://www.biometriccoe.gov/SABER/index.htm, March 2009.

[2] J. Rao and P. Rohatgi, "Can pseudonymity really guarantee privacy?" in *IN PROCEEDINGS OF THE NINTH USENIX SECURITY SYMPOSIUM*, 2000, pp. 85–96.

[3] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 1–29, 2008.

[4] V. Suresh, A. Krishnamurthy, R. Badrinath, and C. Veni Madhavan, "A stylometric study and assessment of machine translators," in *Advances in Intelligent Data Analysis X*, ser. Lecture Notes in Computer Science, J. Gama, E. Bradley, and J. Hollmén, Eds. Springer Berlin / Heidelberg, 2011, vol. 7014, pp. 364–375.

[5] S. Hedegaard and J. G. Simonsen, "Lost in translation: Authorship attribution using frame semantics," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[6] K. Luyckx and W. Daelemans, "Authorship attribution and verification with many authors and limited data," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 513–520.

[7] M. Brennan and R. Greenstadt, "Practical attacks against authorship recognition techniques," *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, 2009.

[8] R. S. Forsyth and D. I. Holmes, "Feature-finding for test classification," *Literary and Linguistic Computing*, vol. 11, no. 4, pp. 163–174, 1996.

[9] A. I. Schein, J. F. Caver, R. J. Honaker, and C. H. Martell, "Author attribution evaluation with novel topic cross-validation," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, October 2010, pp. 206–215.

[10] P. Juola, J. Sofko, and P. Brennan, "A prototype for authorship attribution studies," *Literary and Linguistic Computing*, vol. 21, no. 2, pp. 169–178, 2006.

[11] A. W. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt, "Use fewer instances of the letter "i": Toward writing style anonymization," in *Privacy Enhancing Technologies 12th International Symposium*, 2012.

[12] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Advances in Kernel Methods Support Vector Learning*, vol. 208, no. MSR-TR-98-14, pp. 1–21, 1998.

[13] J. Li, R. Zheng, and H. Chen, "From fingerprint to writeprint," *Commun. ACM*, vol. 49, pp. 76–82, April 2006.

[14] M. Koppel, J. Schler, and K. Zigdon, "Automatically determining an anonymous author's native language," in *Intelligence and Security Informatics*, ser. Lecture Notes in Computer Science, P. Kantor, G. Muresan, F. Roberts, D. Zeng, F.-Y. Wang, H. Chen, and R. Merkle, Eds. Springer Berlin / Heidelberg, 2005, vol. 3495, pp. 41–76.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, pp. 10–18, November 2009.

[16] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online." *IEEE Symposium on Security and Privacy*, 2012.