Lost in Translation: Authorship Attribution using Frame Semantics

Steffen Hedegaard
Department of Computer Science,
University of Copenhagen
Njalsgade 128,
2300 Copenhagen S, Denmark
steffenh@diku.dk

Jakob Grue Simonsen
Department of Computer Science,
University of Copenhagen
Njalsgade 128,
2300 Copenhagen S, Denmark
simonsen@diku.dk

Abstract

We investigate authorship attribution using classifiers based on frame semantics. The purpose is to discover whether adding semantic information to lexical and syntactic methods for authorship attribution will improve them, specifically to address the difficult problem of authorship attribution of translated texts. Our results suggest (i) that frame-based classifiers are usable for author attribution of both translated and untranslated texts; (ii) that frame-based classifiers generally perform worse than the baseline classifiers for untranslated texts, but (iii) perform as well as, or superior to the baseline classifiers on translated texts; (iv) that—contrary to current belief—naïve classifiers based on lexical markers may perform tolerably on translated texts if the combination of author and translator is present in the training set of a classifier.

1 Introduction

Authorship attribution is the following problem: *For a given text, determine the author of said text among a list of candidate authors*. Determining authorship is difficult, and a host of methods have been proposed: As of 1998 Rudman estimated the number of metrics used in such methods to be at least 1000 (Rudman, 1997). For comprehensive recent surveys see e.g. (Juola, 2006; Koppel et al., 2008; Stamatatos, 2009). The process of authorship attribution consists of selecting *markers* (features that provide an indication of the author), and *classifying* a text by assigning it to an author using some appropriate machine learning technique.

1.1 Attribution of translated texts

In contrast to the general authorship attribution problem, the specific problem of attributing translated texts to their original author has received little attention. Conceivably, this is due to the common intuition that the impact of the translator may add enough noise that proper attribution to the original author will be very difficult; for example, in (Arun et al., 2009) it was found that the imprint of the translator was significantly greater than that of the original author. The volume of resources for natural language processing in *English* appears to be much larger than for any other language, and it is thus, conceivably, convenient to use the resources at hand for a translated version of the text, rather than the original.

To appreciate the difficulty of purely lexical or syntactic characterization of authors based on translation, consider the following excerpts from three different translations of the first few paragraphs of Turgenev's Дворянское Гнездо:

Liza "A nest of nobles" Translated by *W. R. Shedden-Ralston*

> A beautiful spring day was drawing to a close. High aloft in the clear sky floated small rosy clouds, which seemed never to drift past, but to be slowly absorbed into the blue depths beyond.
>
> At an open window, in a handsome mansion situated in one of the outlying streets of O., the chief town of the government of that name–it was in the year 1842–there were sitting two ladies, the one about fifty years old, the other an old woman of seventy.

A Nobleman's Nest Translated by *I. F. Hapgood*

> The brilliant, spring day was inclining toward the

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 65–70,
Portland, Oregon, June 19-24, 2011.

evening, tiny rose-tinted cloudlets hung high in the heavens, and seemed not to be floating past, but retreating into the very depths of the azure.

In front of the open window of a handsome house, in one of the outlying streets of O * * * the capital of a Government, sat two women; one fifty years of age, the other seventy years old, and already aged.

A House of Gentlefolk Translated by *C. Garnett*

A bright spring day was fading into evening. High overhead in the clear heavens small rosy clouds seemed hardly to move across the sky but to be sinking into its depths of blue.

In a handsome house in one of the outlying streets of the government town of O—- (it was in the year 1842) two women were sitting at an open window; one was about fifty, the other an old lady of seventy.

As translators express the same semantic content in different ways the syntax and style of different translations of the same text will differ greatly due to the footprint of the translators; this footprint may affect the classification process in different ways depending on the features.

For markers based on language structure such as grammar or function words it is to be expected that the footprint of the translator has such a high impact on the resulting text that attribution to the author may not be possible. However, it is possible that a specific author/translator combination has its own unique footprint discernible from other author/translator combinations: A specific translator may often translate often used phrases in the same way. Ideally, the footprint of the author is (more or less) unaffected by the process of translation, for example if the languages are very similar or the marker is not based solely on lexical or syntactic features.

In contrast to purely lexical or syntactic features, the semantic content is expected to be, roughly, the same in translations and originals. This leads us to hypothesize that a marker based on semantic frames such as found in the FrameNet database (Ruppenhofer et al., 2006), will be largely unaffected by translations, whereas traditional lexical markers will be severely impacted by the footprint of the translator.

The FrameNet project is a database of annotated exemplar frames, their relations to other frames and obligatory as well as optional frame elements for each frame. FrameNet currently numbers approximately 1000 different frames annotated with natural language examples. In this paper, we combine the data from FrameNet with the LTH semantic parser (Johansson and Nugues, 2007), until very recently (Das et al., 2010) the semantic parser with best experimental performance (note that the performance of LTH on our corpora is unknown and may differ from the numbers reported in (Johansson and Nugues, 2007)).

1.2 Related work

The research on authorship attribution is too voluminous to include; see the excellent surveys (Juola, 2006; Koppel et al., 2008; Stamatatos, 2009) for an overview of the plethora of lexical and syntactic markers used. The literature on the use of semantic markers is much scarcer: Gamon (Gamon, 2004) developed a tool for producing semantic dependency graphs and using the resulting information in conjunction with lexical and syntactic markers to improve the accuracy of classification. McCarthy et al. (McCarthy et al., 2006) employed WordNet and latent semantic analysis to lexical features with the purpose of finding semantic similarities between words; it is not clear whether the use of semantic features improved the classification. Argamon et al. (Argamon, 2007) used systemic functional grammars to define a feature set associating single words or phrases with semantic information (an approach reminiscent of frames); Experiments of authorship identification on a corpus of English novels of the 19th century showed that the features could improve the classification results when combined with traditional function word features. Apart from a few studies (Arun et al., 2009; Holmes, 1992; Archer et al., 1997), the problem of attributing translated texts appears to be fairly untouched.

2 Corpus and resource selection

As pointed out in (Luyckx and Daelemans, 2010) the size of data set and number of authors may crucially affect the efficiency of author attribution methods, and evaluation of the method on some standard corpus is essential (Stamatatos, 2009).

Closest to a standard corpus for author attribution is The Federalist Papers (Juola, 2006), originally used by Mosteller and Wallace (Mosteller and Wallace, 1964), and we employ the subset of this

corpus consisting of the 71 undisputed single-author documents as our *Corpus I*.

For translated texts, a mix of authors and translators across authors is needed to ensure that the attribution methods do not attribute to the translator instead of the author. However, there does not appear to be a large corpus of texts publicly available that satisfy this demand.

Based on this, we elected to compile a fresh corpus of translated texts; our *Corpus II* consists of English translations of 19th century Russian romantic literature chosen from Project Gutenberg for which a number of different versions, with different translators existed. The corpus primarily consists of novels, but is slightly polluted by a few collections of short stories and two nonfiction works by Tolstoy due to the necessity of including a reasonable mix of authors and translators. The corpus consists of 30 texts by 4 different authors and 12 different translators of which some have translated several different authors. The texts range in size from 200 (Turgenev: The Rendezvous) to 33000 (Tolstoy: War and Peace) sentences.

The option of splitting the corpus into an artificially larger corpus by sampling sentences for each author and collating these into a large number of new documents was discarded; we deemed that the sampling could inadvertently both smooth differences between the original texts and smooth differences in the translators' footprints. This could have resulted in an inaccurate positive bias in the evaluation results.

3 Experiment design

For both corpora, authorship attribution experiments were performed using six classifiers, each employing a distinct feature set. For each feature set the markers were counted in the text and their relative frequencies calculated. Feature selection was based solely on training data in the inner loop of the cross-validation cycle. Two sets of experiments were performed, each with with $X = 200$ and $X = 400$ features; the size of the feature vector was kept constant across comparison of methods, due to space constraints only results for 400 features are reported. The feature sets were:

Frequent Words (FW): Frequencies in the text of the X most frequent words[1]. Classification with this feature set is used as baseline.

Character N-grams: The X most frequent N-grams for N = $3, 4, 5$.

Frames: The relative frequencies of the X most frequently occurring semantic frames.

Frequent Words and Frames (FWaF): The $X/2$ most frequent features; words and frames resp. combined to a single feature vector of size X.

In order to gauge the impact of translation upon an author's footprint, three different experiments were performed on subsets of Corpus II:

The full corpus of 30 texts [*Corpus IIa*] was used for authorship attribution with an ample mix of authors an translators, several translators having translated texts by more than one author. To ascertain how heavily each marker is influenced by translation we also performed translator attribution on a subset of 11 texts [*Corpus IIb*] with 3 different translators each having translated 3 different authors. If the translator leaves a heavy footprint on the marker, the marker is expected to score better when attributing to translator than to author. Finally, we reduced the corpus to a set of 18 texts [*Corpus IIc*] that only includes unique author/translator combinations to see if each marker could attribute correctly to an author if the translator/author combination was *not* present in the training set.

All classification experiments were conducted using a multi-class winner-takes-all (Duan and Keerthi, 2005) support vector machine (SVM). For cross-validation, all experiments used leave-one-out (i.e. N-fold for N texts in the corpus) validation. All features were scaled to lie in the range $[0, 1]$ before different types of features were combined. In each step of the cross-validation process, the most frequently occurring features were selected from the training data, and to minimize the effect of skewed training data on the results, oversampling with substitution was used on the training data.

[1]The most frequent words, is from a list of word frequencies in the BNC compiled by (Leech et al., 2001)

4 Results and evaluation

We tested our results for statistical significance using McNemar's test (McNemar, 1947) with Yates' correction for continuity (Yates, 1934) against the null hypothesis that the classifier is indistinguishable from a random attribution weighted by the number of author texts in the corpus.

Random Weighted Attribution				
Corpus	I	IIa	IIb	IIc
Accuracy	57.6	28.7	33.9	26.5

Table 1: Accuracy of a random weighted attribution.

FWaF performed better than FW for attribution of author on translated texts. However, the difference failed to be statistically significant.

Results of the experiments are reported in the table below. For each corpus results are given for experiments with 400 features. We report macro[2] precision/recall, and the corresponding F1 and accuracy scores; the best scoring result in each row is shown in **boldface**. For each corpus the bottom row indicates whether each classifier is significantly discernible from a weighted random attribution.

400 Features							
Corpus	Measure	FW	3-grams	4-grams	5-grams	Frames	FWaF
I	precision	96.4	97.0	97.0	**99.4**	80.7	92.0
	recall	90.3	97.0	91.0	**97.6**	66.8	93.3
	F1	93.3	97.0	93.9	**98.5**	73.1	92.7
	Accuracy	95.8	97.2	97.2	**98.6**	80.3	93.0
	p<0.05:	✓	✓	✓	✓	✓	✓
IIa	precision	63.8	61.9	59.1	57.9	**82.7**	81.9
	recall	66.4	60.4	60.4	60.4	70.8	**80.8**
	F1	65.1	61.1	59.7	59.1	76.3	**81.3**
	Accuracy	80.0	73.3	73.3	73.3	76.7	**90.0**
	p<0.05:	✓	✓	✓	✓	✓	✓
IIb	precision	**91.7**	47.2	47.2	38.9	70.0	70.0
	recall	**91.7**	58.3	58.3	50.0	63.9	63.9
	F1	**91.7**	52.2	52.2	43.8	66.8	66.8
	Accuracy	**90.9**	63.6	63.6	54.5	63.6	63.6
	p<0.05:	✓	⁒	⁒	⁒	⁒	⁒
IIc	precision	42.9	43.8	42.4	51.0	60.1	**75.0**
	recall	52.1	42.1	42.1	50.4	59.6	**75.0**
	F1	47.0	42.9	42.2	50.7	59.8	**75.0**
	Accuracy	55.6	50.0	44.4	55.6	61.1	**72.2**
	p<0.05:	⁒	⁒	⁒	⁒	⁒	✓

Table 2: Authorship attribution results

[2]each author is given equal weight, regardless of the number of documents

4.1 Corpus I: The Federalist Papers

For the Federalist Papers the traditional authorship attribution markers all lie in the 95+ range in accuracy as expected. However, the frame-based markers achieved statistically significant results, and can hence be used for authorship attribution on untranslated documents (but performs worse than the baseline). FWaF did not result in an improvement over FW.

4.2 Corpus II: Attribution of translated texts

For Corpus IIa–the entire corpus of translated texts– all methods achieve results significantly better than random, and FWaF is the best-scoring method, followed by FW.

The results for Corpus IIb (three authors, three translators) clearly suggest that the footprint of the translator is evident in the translated texts, and that the FW (function word) classifier is particularly sensitive to the footprint. In fact, FW was the only one achieving a significant result over random assignment, giving an indication that this marker may be particularly vulnerable to translator influence when attempting to attribute authors.

For Corpus IIc (unique author/translator combinations) decreased performance of all methods is evident. Some of this can be attributed to a smaller (training) corpus, but we also suspect the lack of several instances of the same author/translator combinations in the corpus.

Observe that the FWaF classifier is the only classifier with significantly better performance than weighted random assignment, and outperforms the other methods. Frames alone also outperform traditional markers, albeit not by much.

The experiments on the collected corpora strongly suggest the feasibility of using Frames as markers for authorship attribution, in particular in combination with traditional lexical approaches.

Our inability to obtain demonstrably significant improvement of FWaF over the approach based on Frequent Words is likely an artifact of the fairly small corpus we employ. *However*, computation of significance is generally woefully absent from studies of automated author attribution, so it is conceivable that the apparent improvement shown in many such studies fail to be statistically significant under

closer scrutiny (note that the exact tests to employ for statistical significance in information retrieval–including text categorization–is a subject of contention (Smucker et al., 2007)).

5 Conclusions, caveats, and future work

We have investigated the use of semantic frames as markers for author attribution and tested their applicability to attribution of translated texts. Our results show that frames are potentially useful, especially so for translated texts, and suggest that a combined method of frequent words and frames can outperform methods based solely on traditional markers, on translated texts. For attribution of untranslated texts and attribution to translator traditional markers such as frequent words and n-grams are still to be preferred.

Our test corpora consist of a limited number of authors, from a limited time period, with translators from a similar limited time period and cultural context. Furthermore, our translations are all from a single language. Thus, further work is needed before firm conclusions regarding the general applicability of the methods can be made.

It is well known that effectiveness of authorship markers may be influenced by topics (Stein et al., 2007; Schein et al., 2010); while we have endeavored to design our corpora to minimize such influence, we do not currently know the quantitative impact on topicality on the attribution methods in this paper. Furthermore, traditional investigations of authorship attribution have focused on the case of attributing texts among a small ($N < 10$) class of authors at the time, albeit with recent, notable exceptions (Luyckx and Daelemans, 2010; Koppel et al., 2010). We test our methods on similarly restricted sets of authors; the scalability of the methods to larger numbers of authors is currently unknown. Combining several classification methods into an ensemble method may yield improvements in precision (Raghavan et al., 2010); it would be interesting to see whether a classifier using frames yields significant improvements in ensemble with other methods. Finally, the distribution of frames in texts is distinctly different from the distribution of words: While there are function *words*, there are no 'function frames', and certain frames that are common in a corpus may fail to occur in the training material of a given author; it is thus conceivable that smoothing would improve classification by frames more than by words or N-grams.

References

John B. Archer, John L. Hilton, and G. Bruce Schaalje. 1997. Comparative power of three author-attribution techniques for differentiating authors. *Journal of Book of Mormon Studies*, 6(1):47–63.

Shlomo Argamon. 2007. Interpreting Burrows' Delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147.

R. Arun, V. Suresh, and C. E. Veni Madhaven. 2009. Stopword graphs and authorship attribution in text corpora. In *Proceedings of the 3rd IEEE International Conference on Semantic Computing (ICSC 2009)*, pages 192–196, Berkeley, CA, USA, sep. IEEE Computer Society Press.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Proceedings of the North American Chapter of the Association for Compututional Linguistics Human Language Technologies Conference (NAACL HLT '10)*.

Kai-Bo Duan and S. Sathiya Keerthi. 2005. Which is the best multiclass svm method? an empirical study. In *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, pages 278–285.

Michael Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, pages 611–617.

David I. Holmes. 1992. A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society, Series A*, 155(1):91–120.

Richard Johansson and Pierre Nugues. 2007. Semantic structure extraction using nonprojective dependency trees. In *Proceedings of SemEval-2007*, Prague, Czech Republic, June 23-24.

Patrick Juola. 2006. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2008. Computational methods for authorship attribution. *Journal of the American Society for Information Sciences and Technology*, 60(1):9–25.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2010. Authorship attribution in the wild. *Language Resources and Evaluation*, pages 1–12. 10.1007/s10579-009-9111-2.

Geoffrey Leech, Paul Rayson, and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Longman, London.

Kim Luyckx and Walter Daelemans. 2010. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*. To appear.

Philip M. McCarthy, Gwyneth A. Lewis, David F. Dufty, and Danielle S. McNamara. 2006. Analyzing writing styles with coh-metrix. In *Proceedings of the International Conference of the Florida Artificial Intelligence Research Society*, pages 764–769.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157.

Frederick Mosteller and David L. Wallace. 1964. Inference and Disputed Authorship: The Federalist. Springer-Verlag, New York. 2nd Edition appeared in 1984 and was called *Applied Bayesian and Classical Inference*.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42. Association for Computational Linguistics.

Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365.

Joseph Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. The Framenet Project.

Andrew I. Schein, Johnnie F. Caver, Randale J. Honaker, and Craig H. Martell. 2010. Author attribution evaluation with novel topic cross-validation. In *Proceedings of the 2010 International Conference on Knowledge Discovery and Information Retrieval (KDIR '10)*.

Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 623–632, New York, NY, USA. ACM.

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

Benno Stein, Moshe Koppel, and Efstathios Stamatatos, editors. 2007. *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection, PAN 2007, Amsterdam, Netherlands, July 27, 2007*, volume 276 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Frank Yates. 1934. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):pp. 217–235.