doi:10.5128/ERYa10.16

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 10, 259–267

AUTHORSHIP VERIFICATION OF OPINION PIECES IN ESTONIAN

Timo Petmanson

Abstract. Authorship verification is an important subproblem in authorship attribution and plagiarism detection tasks. We present a novel approach for extracting stylistic features unique to individual authors. We use the correlations of important textual features as a way to learn the style. The goal of our proposed method is to answer the following question: given a set of documents known to be written by the same person and an unknown document, is the unknown document also written by that individual. We present the first study of this problem conducted on opinion pieces written in Estonian. Our method achieves 74% precision, which is comparable with current state-of-the-art systems tested in other languages, whereas the recall level is still something to be improved on.

Keywords: natural language processing, text analysis, linguistic expertise, machine learning, pattern mining, feature correlations, Estonian

1. Introduction

The question of whether any two documents are written by the same author was proposed as a fundamental aspect of every authorship attribution problem by Koppel, Winter (2014). Authorship verification arises in many case studies such as police investigations requiring detection of the ownership of defamatory letters and ransom notes (De Vel et al. 2001, Abbasi, Chen 2005). It could also be used to learn the general profile of a group of persons such as sexual predators (Inches, Crestani 2012).

We address an extended version of the problem: given several documents known to be written by the same author, is that particular person the author of one or more unknown documents. This problem has been also addressed in PAN Workshop and Competition: Uncovering Plagiarism, Authorship and Social Software Misuse (Argamon, Juola 2011, Juola, Stamatos 2013). Although authorship verification is

also important in other domains such as source code analysis (Frantzeskou et al. 2004), we concentrate on texts.

What makes this problem particularly difficult is that we are given only documents from a single author, not from a closed set of candidates. Although authorship attribution and identification are relatively well-studied areas, most of the literature addresses case studies with a closed set of authors, which might not be sufficient in real-world scenarios (Koppel et al. 2009). The task formalized in open-set fashion has gained more interest recently.

There have been a few studies related to authorship verification in Estonian, such as plagiarism detection in dictionaries (Langemets, Voll 2008). However, this study is the first addressing the issue directly, setting the baseline for future studies.

2. Data description

Our Corpus of Opinion Pieces of Estonian covers 1474 news articles from 318 authors, where 295 of them have ten articles or less. Figure 1 depicts a more detailed distribution of the number of documents per author. As many as 123 authors have only two articles in our corpus. We did not use authors that have only a single article, because we need at least one document for training and one for evaluation. The corpus is a subset of opinion pieces (editorials, columns, etc.) published by Postimees Online news portal from January 2008 till September 2013.

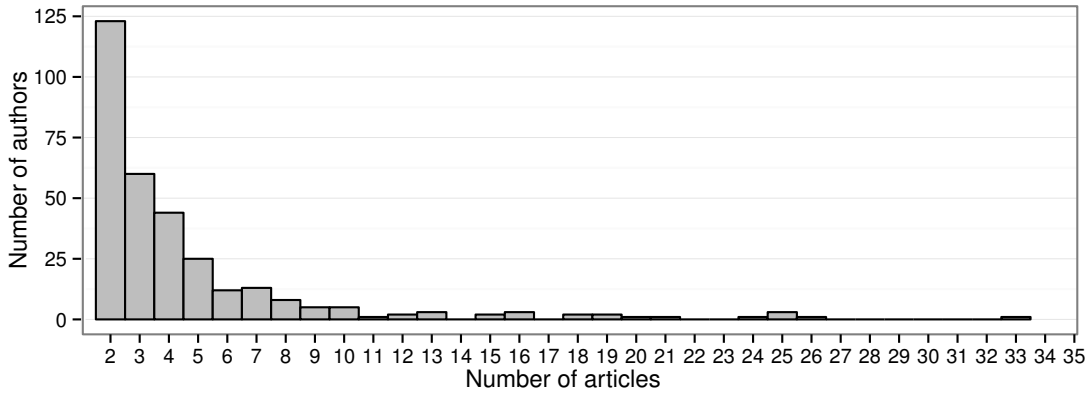

Figure 1. Histogram depicting the number of documents per author in the corpus

Our computational methods, which we describe in the following sections, require a reference corpus. For this task, we decided to use a subset of opinion pieces from the Balanced Corpus of Estonian[1] which were not written by journalists presented in our main corpus.

Authorship identification of news articles is particularly difficult as the articles are edited before publishing. The style and language use is more normalized compared to more informal texts such as emails and blogs. This makes it harder to exploit certain features like usage of emoticons, number of grammar errors, etc. which are often representative of people of a certain age and gender (Koppel, Argamon, Shimoni 2002, Stamatatos 2009). Although news articles hide many stylistic aspects the authors would use informally, the news corpus makes an interesting example of

[1] The Balanced Corpus of Estonian http://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en (31.1.2014).

detecting less obvious patterns. This might be important in cases of someone trying to impersonate somebody else or intentionally hiding his/her own writing style.

3. Feature extraction

Common features used in authorship verification are traditional in natural language processing: word and character n-grams, word prefixes and suffixes of various sizes, exploiting external sources such as Wordnet and Wikipedia (Houvardas, Stamatatos 2006, Brocardo et al. 2013). We experimented with various combinations of features and decided to use the most easily interpretable ones. We used the following formal and lexical features: is the word uppercase, does the word start with an uppercase letter, does the word contain a digit. From morphological features we considered word lemma, part-of-speech (POS), case and verb type. Note that not all possible features exist for all words. For example, verbs do not have cases.

Another reason we did not include traditional n-grams is that typical information about the words in our corpus is already encoded by the lemma and other morphological features. N-grams would be valuable for capturing the use of emoticons, excessive punctuation and significant amount of grammar and typing errors. This would be important in more informal texts like e-mails, but not in our corpus.

Table 1. Example of important features used in the study

word	lemma	POS	case	verb type	uppercase
Usjas	usjas	adjective	nominative	–	yes
kaslane	kaslane	noun	nominative	–	no
jookseb	jooksma	verb	–	b	no
künklikul	künklik	adjective	adessive	–	no
maastikul.	maastik	noun	adessive	–	no

As an example, consider the following sentence: “Usjas kaslane jookseb künklikul maastikul” (‘The sinuous feline is running on a hilly terrain’). Table 1 shows the extracted features for the sentence. The morphological information was extracted using the morphological analyzer *t3mesta* (Kaalep 1997, Kaalep, Vaino 1998).

3.1. Pattern mining

Simple features described in Table 1 are good for many classic tasks like clustering the documents by topic. However, we might get more descriptive stylistic features if we combine several of them. For example, an author may have a unique way of ordering certain types of words in some phrases. Some of this information can be captured, if we encode each feature as a pattern

(feature, offset, value)

where *offset* determines the relative position of the feature/value combination. For instance, a pattern *(case, -1, nominative)* would say that the case of the previous word is nominative. Using the offset makes sense, when we want to combine two

simpler patterns. For example, a composite pattern identifying all words in the partitive case followed by a verb is *(case, 0, partitive) & (pos-tag, 1, verb)*. The maximal absolute offset was two.

Although we extracted one set of patterns from all words, we also extracted sets of them specifically related to noun, verb and adjective phrases. We achieved this by separately considering contexts of nouns, verbs and adjectives with a radius of two words. A final set of patterns were extracted from contexts surrounding punctuation. Punctuation patterns may give insight to unique formatting preferences.

We used the PatNLP library (Petmanson, Laur 2012) to mine the frequent patterns for each author. We instructed the library to detect patterns covering at least five percent of the tokens. Smaller thresholds were not computationally feasible and larger thresholds yielded too generic patterns.

Not all the frequent patterns describe the writing style of the author, but can be just specific to the Estonian language or the news article style in general. Such patterns are not discriminative enough to capture the style of a journalist. To tell the relevance of a pattern, we estimated the p-value using 500 randomly selected articles from the Balanced Corpus of Estonian as a reference. The p-value was computed as the percentage of documents where the pattern was at least as frequent as in the articles of the author. The smaller the p-value, the more relevant and surprising the pattern is. We kept only patterns having p-value less than 2.5 percent. The threshold was chosen such that it would yield about a few hundred statistically significant patterns.

3.2. Classification model

As previously explained, we mined the patterns separately for five groups of tokens: all words, nouns, verbs, adjectives and punctuation. The simplest way to use them as features for a machine learning classifier is to encode them as a vector of frequencies – a single vector for a single document. However, such patterns only capture document level co-occurrence of features. Information on whether two frequent and statistically significant patterns are never used together in the same phrase or sentence is discarded. However, this kind of knowledge might be useful for classification.

To address the issue, we encoded the matches of the patterns as bitvectors, where a *true* bit indicates a match. Next, we computed the *Matthews correlation coefficient* (MCC) between all bitvectors. Note that we used MCC instead of alternatives such as the *Pearson correlation coefficient* because the matches are binary events. The Pearson correlation assumes that the events follow a Gaussian distribution, whereas MCC is specifically designed for bitvectors. Now, by encoding the correlations as a single vector, we can use it as an input for the classifier.

As our problem statement assumed an open set of authors, we can only learn from known documents, i.e. positive examples. Such a problem is known as a one-class learning problem in machine learning and has many solutions. In this work, we used a 1-class *support vector machine* (Cortes, Vapnik 1995) to perform the classification task. We employed the Python *scikit-learn* library (Pedregosa et al. 2011) implementation using the RBF kernel with default parameters.

4. Results

To evaluate our proposed method, we created training and testing collections for each author in the Corpus of Opinion Pieces of Estonian (see Chp. 2). In order to build the model, we could only use the documents written by a single person. As our corpus also contained authors with only two articles, the testing corpora were designed so that they contained one randomly selected document from the same author and another from a different author. The training set contained the rest of the known articles by the author. The total number of testing documents in all collections was 636 with 318 articles for both positive and negative examples respectively.

We built a separate model for each author using the respective training set and evaluated that particular model on the testing set of the same author. Next, we computed the overall number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) hits (see Table 2). Using these counts, we computed relevant statistical measures for the method (see Table 3).

We see that our method has an acceptable precision of 74%, low recall of 14%, which combined as an F1-score is 23%. Current state-of-the art authorship identification systems achieve an F1-score of 75% on the PAN'13 corpus containing various documents in Spanish, English and Greek (Juola, Stamatos 2013). We need to note that the evaluation setup used in (Juola, Stamatos 2013) also included correctly classified negative examples in precision and recall computation. The reason was that they did not require all the questions to be answered. Thus, resulting precision, recall and F1-scores were all equal if a system answered all the questions. But we cannot directly compare the performances, since the datasets on which the algorithms were evaluated are different.

Table 2. Number of true positive (TP), true negative (TN), false positive (FP), false negative (FN) hits for authors with more than 10 training documents (N > 10), with less or equal to ten documents (N ≤ 10) and all authors

Datasets	TP	TN	FP	FN
N ≤ 10	10	286	10	261
N > 10	35	16	6	12
All	45	302	16	273

Table 3. Precision, recall, accuracy and F1-scores computed from Table 2

Metric	N ≤ 10	N > 10	All
precision	78%	63%	74%
recall	12%	45%	14%
accuracy	54%	59%	55%
F1-score	20%	52%	23%

Our method is very conservative for authors having ten training articles or less. We achieve good precision of 78%, but low recall. Authors with more than ten training documents achieve better recall of 45%, albeit worse precision of 63%.

As the topic and word usage in documents are very important features, we decided to study them and see if we could find any link between misclassified

examples. In Table 4, we have applied the Latent Semantic Indexing (Deerwester et al. 1990) method to extract the most influential keywords of the central topic in our document collections.

Table 4. Main keywords identifying the central topic in all documents, true positive and false positive examples

Document collection	LSI main topic keywords
All documents	0.5*"*eesti*", 0.3*"*aasta*", 0.2*"*riik*", 0.2*"*inimene*", 0.1*"*laps*"
True positive documents	0.4*"*eesti*", 0.3*"*aasta*", 0.2*"*inimene*", 0.2*"*riik*", 0.2*"*laps*"
False positive documents	0.5*"*eesti*", 0.2*"*aasta*", 0.2*"*riik*", 0.2*"*inimene*", 0.1*"*euroopa*"
False negative documents	-0.4*"*laps*", -0.3*"*eesti*", -0.2*"*aasta*", -0.2*"*inimene*", -0.2*"*riik*"

We fitted the LSI model separately on all true positive, true negative, false positive and false negative documents and extracted the most influential keywords. Coefficients show the weights of the keywords attributing to the central topic. A negative coefficient means that a particular keyword is statistically underrepresented in the document collection.

We see that for all documents the main keywords describing the central topic are, in order of importance, *eesti* ('Estonia', 'Estonian'), *aasta* ('year'), *riik* ('state'), *inimene* ('person') and *laps* ('child'). Almost the same keywords with similar coefficients are also present in true and false positive document corpora. Although the number of false positives was relatively small, it seems they were misclassified as the content was very similar, albeit by a different author. False negative documents seem in general not to be about the main topic in other collections (note the negative coefficients of main keywords in Table 4), indicating' that the profile learned by the classifier included too many topic-related features.

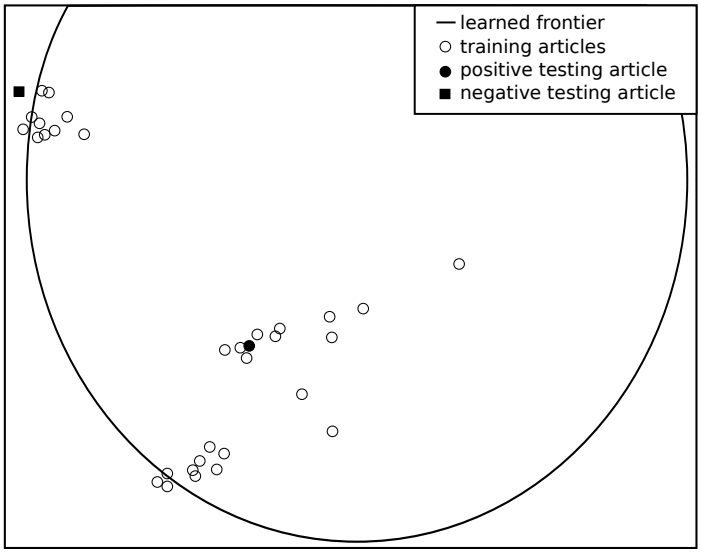

Figure 2. Frontier learned by support vector machine for columnist Ahto Lobjakas. The full coordinate space is transformed to 2D-coordinate space by applying principal component analysis. Everything inside the frontier would be classified as from the same author and everything outside as from a different author

Another possible reason for misclassification is that some authors do not in fact have a single profile or style, but may have several. In Figure 2, we have plotted document representations from a single training and testing corpora pair. We see at least three distinctive clusters of documents, which correspond to different styles or topics of the author. Although in this particular case both unknown documents were classified correctly, the model representation could have been better. Instead of a single large frontier, we might have tried to learn three smaller frontiers instead, each capturing the specifics of representative clusters. How to exploit this will be of particular interest of our future studies.

5. Summary

In this work, we described a novel method using pattern mining and feature correlations for the task of authorship verification. We achieved acceptable precision of 74%, albeit not very satisfying recall. The main reason for misclassification seemed to be the small set of training samples and the large impact of the topic on the final authorship model. Although the topic of the documents can be helpful in the case of niche authors, our future studies should concentrate on how to separate topics from the unique stylistic features of the authors.

References

Abbasi, Ahmed; Chen, Hsinchun 2005. Applying authorship analysis to extremist-group web forum messages. – Intelligent Systems, IEEE, 20 (5), 67–75. http://dx.doi.org/10.1109/MIS.2005.81

Argamon, Shlomo; Juola, Patrick 2011. Overview of the International Authorship Identification Competition at PAN-2011. – V. Petras, P. Forner, P. D. Clough (Eds.). CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands. http://www.informatik.uni-trier.de/~ley/db/conf/clef/clef2011w.html (1.2.2014).

Brocardo, Marcelo Luiz; Traore, Issa; Saad, Sherif; Woungang, Isaac 2013. Authorship verification for short messages using stylometry. – Proceedings of the IEEE 2013 International Conference on Computer, Information and Telecommunication Systems (CITS), 1–6. http://dx.doi.org/10.1109/CITS.2013.6705711

Cortes, Corinna; Vapnik, Vladimir 1995. Support-vector networks. – Machine Learning, 20 (3), 273–297. http://dx.doi.org/10.1007/BF00994018

De Vel, Olivier; Anderson, Alison; Corney, Malcolm; Mohay, George 2001. Mining e-mail content for author identification forensics. – ACM Sigmod Record, 30 (4), 55–64. http://dx.doi.org/10.1145/604264.604272

Deerwester, Scott; Dumais, Susan; Landauer, Thomas; Furnas, George; Harshman, Richard 1990. Indexing by latent semantic analysis. – Journal of the American Society for Information Science (JASIS), 41 (6), 391–407. http://www.informatik.uni-trier.de/~ley/db/journals/jasis/jasis41.html (1.2.2014).

Frantzeskou, Georgia; Gritzalis, Stefanos; MacDonell, Stephen 2004. Source code authorship analysis for supporting the cybercrime investigation process. – Proceedings of the 1st International Conference on e-business and Telecommunications Networks (ICETE04), Setúbal, Portugal, 85–92. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.104.6476 (1.2.2014).

Houvardas, John; Stamatatos, Efstathios 2006. N-gram feature selection for authorship identification. – Jérôme Euzenat, John Domingue (Eds.). Artificial Intelligence: Methodology, Systems, and Applications. 12th International Conference, AIMSA 2006, Varna, Bulgaria, September 12-15, 2006. Proceedings. Lecture Notes in Computer Science 4183. Berlin, Heidelberg: Springer, 77–86.

Inches, Giacomol; Crestani, Fabio 2012. Overview of the International Sexual Predator Identification Competition at PAN-2012. – P. Forner, J. Karlgren, C. Womser-Hacker (Eds.). CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20. http://dblp.uni-trier.de/db/conf/clef/clef2012w.html#InchesC12 (1.2.2014).

Juola, Patrick; Stamatos, Efsthathios 2013. Overview of the Author Identification Task at PAN 2013. – P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.). Information Access Evaluation. Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings. Springer. http://www.clef-initiative.eu/documents/71612/3095ffc3-376b-40eb-af10-8251c5f107f6 (29.9.2013).

Kaalep, Heiki-Jaan 1997. An Estonian morphological analyser and the impact of a corpus on its development. – Computers and the Humanities, 31 (2), 115–133. http://dx.doi.org/10.1023/A:1000668108369

Kaalep, Heiki-Jaan; Vaino, Tarmo 1998. Kas vale meetodiga õiged tulemused? Eesti keele morfoloogiline ühestamine statistika abil. – Keel ja Kirjandus, 1, 30–38.

Koppel, Moshe; Argamon, Shlomo; Shimoni, Anat Rachel 2002. Automatically categorizing written texts by author gender. – Literary and Linguistic Computing, 17 (4), 401–412. http://dx.doi.org/10.1093/llc/17.4.401

Koppel, Moshe; Schler, Jonathan; Argamon, Shlomo 2009. Computational methods in authorship attribution. – Journal of the American Society for information Science and Technology, 60 (1), 9–26. http://dx.doi.org/10.1002/asi.20961

Koppel, Moshe; Winter, Yaron 2014. Determining if Two Documents are by the Same Author. – Journal of the American Society for Information Science and Technology. Journal of the Association for Information Science and Technology, 65 (1), 178–187. http://dx.doi.org/10.1002/asi.22954

Langemets, Margit; Voll, Piret 2008. Sõnaraamatu kohtulingvistiline analüüs: Eesti pretsedent. [Linguistic forensic analysis of a dictionary: an Estonian precedent.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 4, 67–86. http://dx.doi.org/10.5128/ERYa4.05

Pedregosa, Fabian; Varoquaux, Gaël; Gramfort, Alexandre; Michel, Vincent; Thirion, Bertrand; Grisel, Olivier; Blondel, Mathieu; Prettenhofer, Peter; Weiss, Ron; Dubourg, Vincent; Vanderplas, Jake; Passos, Alexandre; Cornapeau, David 2011. Scikit-learn: Machine learning in Python. – The Journal of Machine Learning Research, 12, 2825–2830.

Petmanson, Timo; Laur, Sven 2012. Pattern based fact extraction from Estonian texts. National Programme for Estonian Language Technology. Project Report. University of Tartu. February 17, 2012. http://www.keeletehnoloogia.ee/ekt-projektid/mallipohine-faktituletus-tekstikorpustest/projekti-tulemusi-kajastavad-uurimisraportid/esimene-vaheraport (1.2.2014).

Stamatatos, Efstathios 2009. A survey of modern authorship attribution methods. – Journal of the American Society for Information Science and Technology, 60 (3), 538–556. http://dx.doi.org/10.1002/asi.21001

AUTORITUVASTUS EESTIKEELSETES ARVAMUSARTIKLITES

Timo Petmanson
Tartu Ülikool

Autorituvastus on üks plagiarismituvastuse olulisi alamprobleeme. Käesolevas töös pakume välja autorituvastuse algoritmi eestikeelsete arvamusartiklite jaoks. Ajakirjaniku tuvastamiseks otsime tema tekstidest mustreid, mis kirjeldavad kõige paremini just talle omaseid lause- ja sõnakonstruktsioone.

Autorituvastuse jaoks on aegade jooksul pakutud välja mitmeid lahendusi, kuid reeglina on alati eeldatud kindlat kandidaatautorite hulka. Käesolevas töös vaatame selle probleemi keerukamat kuju, kus võimalike autorite hulk on täpsustamata ning tuleb vastata vaid küsimusele, kas etteantud tekst on konkreetse autori kirjutatud või mitte. Selle ülesande piisavalt hea lahendus omab mitmeid praktilisi rakendusi, nagu näiteks anonüümsete laimukirjade autorite tuvastus.

Siinne artikkel on esimene, mis käsitleb probleemi eestikeelsetel tekstidel. Kirjeldatud meetod saavutab Postimees Online korpusel 74% täpsuse. Selline täpsus on võrreldav parimate süsteemidega teiste keelte jaoks, kuid samas on 14% saagis siiski liiga madal.

Võtmesõnad: keeletöötlus, tekstianalüüs, keeleekspertiis, masinõpe, mustrikaeve, tunnuste korrelatsioon, eesti keel

Timo Petmanson is a second year PhD student at the Institute of Computer Science of the University of Tartu and works on topics related to machine learning and language technology.
Liivi 2, 50409 Tartu, Estonia
timo_p@ut.ee