Chapter 29

Keystroke dynamics for authorship attribution

Barbara Plank
University of Groningen

We examine to what extent information from keystroke dynamics reflects individual style for authorship attribution. We compare models that use keystroke dynamics to more traditional authorship attribution methods. Our results show that biometric features are more predictive of authorship than stylometric features.

Authorship attribution is the task of identifying the author of a text. It can be viewed as a special form of text classification in the field of stylometry, which more broadly speaking includes the identification of author traits (like identity, or gender, age, personality etc).

As noted by Nerbonne (2007): "A key question in authorship attribution has been to determine what sorts of *evidence* might bear on determining authorship." Traditionally, authorship studies focused on finding evidence in the *text* produced by authors, and examined, e.g., high-frequency elements.

However, as people produce text, they unconsciously produce loads of cognitive by-product. Can we use such meta-data as additional evidence of authorship? Examples of cognitive processing data include brain activations, gaze pattern, or keystroke dynamics. In this paper we focus on the latter. *Keystroke dynamics* concerns a user's typing pattern and keystroke logs are the recordings of a user's typing dynamics. When a person types on a keyboard, the latencies between successive keystrokes and their duration reflect the presumably unique typing behavior of a person. Assuming access to keystroke logs, to what extent are they informative for authorship attribution, i.e., do they help identifying the author of a piece of text?

Keystroke logs are studied mostly in cognitive writing and translation process research to gain insights into the cognitive load involved in the writing process. Only very recently this source has been explored as information in natural language processing (NLP), in particular for shallow syntactic parsing (Plank 2016). Keystroke logs have been used in computer security for user verification, however, combining keystroke biometrics with traditional stylometry metrics has not yet been proven successful (Stewart et al. 2011). In this paper we examine to what extent keystroke dynamics are informative for authorship attribution. We compare them to more traditional stylometry features, and investigate various ways to combine them.

1 Keystroke dynamics

Keystroke dynamics provide a complementary view on a user's style beyond the linguistic signal.

A major scientific interest in keystroke dynamics arose in writing research, where it has developed into a promising non-intrusive method for studying cognitive processes involved in writing (Sullivan & Lindgren 2006; Nottbusch, Weingarten & Sahel 2007; Wengelin 2006; van Waes, Leijten & van Weijen 2009; Baaijen, Galbraith & de Glopper 2012). In these studies time measurements—pauses, bursts and revisions—are studied as traces of the recursive nature of the writing process. *Bursts* are defined as consecutive chunks of text produced and defined by a 2000ms time of inactivity (Wengelin 2006).

Keystroke logs have the distinct advantage over other cognitive modalities like eye tracking or brain scanning that they are readily available and can be harvested easily. They do not rely on special equipment beyond a keyboard. Moreover, they are non-intrusive, inexpensive, and have the potential to offer continuous adaptation to specific users. Imagine integrating keystroke logging into (online) text processing tools.

In its raw form, keystroke logs contain information on which key was pressed for how long (key, time press, time release). Research on keystroke dynamics typically considers a number of timing metrics, such as *holding time* and *time press* and *time release* between keystrokes, e.g., p in Figure 1, inspired by the figure in (Goodkind & Rosenberg 2015). An example of raw keystroke log data is shown in Table 1.

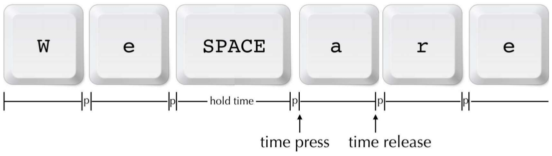

Figure 1: Keystroke logs illustrated: p are pauses between keystrokes.

Raw keystroke log data can be used to calculate keystroke pause durations, such as pre-word pauses. However, if we examine the literature we find different ways to define the duration of pauses. Stewart et al. (2011) and Goodkind & Rosenberg (2015) use the difference between release time of the previous key and the timepress of the current key to calculate keystroke (or pre-word durations). In contrast, writing research (Wengelin 2006; van Waes, Leijten & van Weijen 2009; Baaijen, Galbraith & de Glopper 2012) defines pauses as the start time of a keystroke until the start time for the next keystroke. In this paper we follow user authentication studies (Stewart et al. 2011) and use the former definition of pause duration.

Table 1: Example of the raw keystroke logging data.

user	session	timepress	timerelease	keycode	keyname
1	1	1304433167859	1304433168307	16	shift
1	1	1304433168227	1304433168371	67	c
1	1	1304433168291	1304433168451	79	o
1	1	1304433170051	1304433170179	69	e
1	1	1304433170451	1304433170531	70	f
1	1	1304433170579	1304433170675	70	f
1	1	1304433170675	1304433170851	73	i
1	1	1304433171171	1304433171299	67	c
1	1	1304433172179	1304433172275	8	backspace

A challenge when using keystroke log data is that the typing behavior of users typically differ. For instance, Figure 2 plots the distribution of keystroke durations for two different users. Keystroke logs are presumably idiosyncratic. In fact, they were successfully used for author verification in computer security research (Stewart et al. 2011; Monaco et al. 2013; Locklear et al. 2014). In this paper we study how predictive biometric features are for authorship, as compared to more traditional features obtained from the text alone, and whether combining the two sources aids authorship attribution.

2 Experiments

Given a dataset with keystroke logs from 38 authors, the aim of our experiments is to classify who of the authors wrote a piece of text.

2.1 Dataset

The keystroke logging data stems from students taking an actual test on spreadsheet modeling in a university course (Stewart et al. 2011). The dataset was collected during an exam, and as such represents free-text input. The dataset contains data from 38 users for several sessions.[1] We take the first two sessions as development data (resulting in 76 instances), session 3-5 as test section (114 instances), and the remaining session as training sections (total 856 instances).

2.2 Setup

As classification system we use Support Vector Machines (SVM), implemented in `sk-learn`.[2] For all experiments we use the same hyperparameters, i.e., SVM with

[1] Following Stewart et al. (2011) some users were discarded due to issues with logging.

[2] The code for our experiments is available at: `https://github.com/bplank/festschrift`.

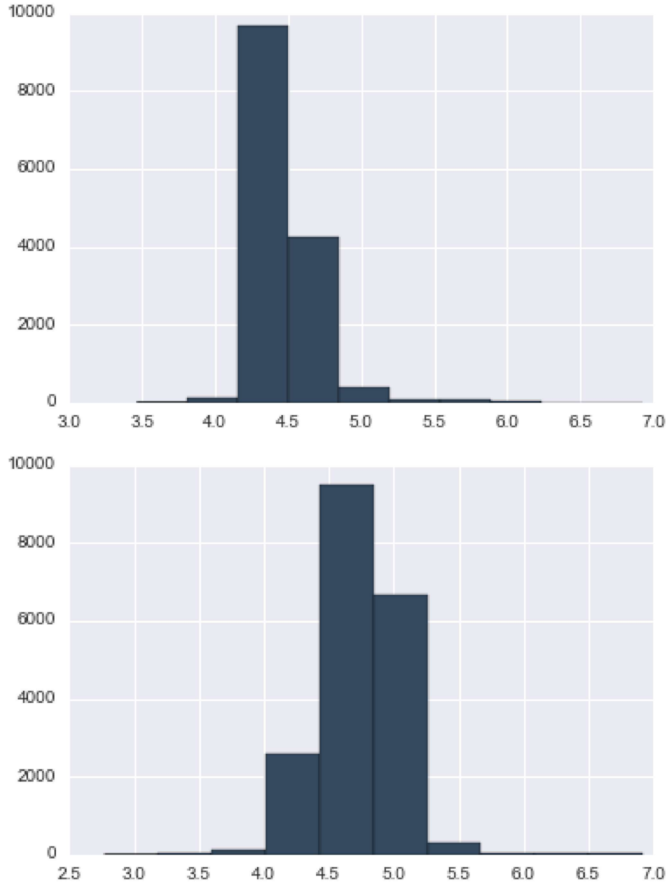

Figure 2: Distribution of pauses for two users (plotted in log space): user 3 (top), user 20 (bottom).

default C and a linear kernel. SVMs were chosen in preliminary experiments as they outperformed alternative approaches (logistic regression, naive Bayes).

2.3 Features

We use 218 biometric features following Stewart et al. (2011), who in turn follow Tappert et al. (2010). These biometric features include duration features (mean and standard deviation) and are grouped roughly into: duration features of individual letters (which we later refer to as keystroke basic), and transition features between letters or groups of letters, between letters and non-letters and overall percentage features.

We use the freely available feature extractor[3] and test two configurations: using only letter durations (52 keystroke basic features) and all duration features (keystrokes extended, 218 features).

For the textual features, we extracted the final text from the keystroke logging data (using revisions to alter the text to obtain the final output). We then use commonly used authorship attribution features, binary indicator features for character n-grams and word n-grams. We also evaluated only pronouns, however, that resulted in worse performance, thus we do not further elaborate on it. In addition, we examine *word embedding* features estimated from a large English Wikipedia dump (Al-Rfou, Perozzi & Skiena 2013). We represent each text as the mean average activation over all word embeddings (Collobert et al. 2011), resulting in 64 features. In contrast to the previous *sparse* n-gram feature representations, this represents adding a *dense* feature vector that represents the text, and is more similar to the standardized keystroke features (both in terms of value and number of features).

3 Results

The results of training a classifier to predict the identify on an author are given in Table 2. A random baseline obtains an accuracy rate of only 2% on this dataset (38 authors). The stylistic features based on the text obtain an accuracy of around 28-37%.

Table 2: Accuracy for authorship attribution (38 authors), comparison of stylometry features (word and character n-grams) versus biometric stylometry (keystroke dynamics) and combined (embeds: word embeddings).

Features	num features	Dev	Test
Stylistic:			
character 3grams	8.3k	23.68	28.07
character 2+3grams	9.8k	25.00	31.58
word unigrams	8.9k	27.64	30.70
word unigrams +char 2+3grams	18.7k	25.00	37.72
Biometrics:			
keystrokes (basic)	52	72.37	71.05
keystrokes (extended)	218	81.58	77.19
Combined:			
keystrokes (basic)+word unigrams	8.9k	55.26	50.88
keystrokes (ext.)+word unigrams	9.1k	65.79	67.65
keystroke (basic)+embeds	116	73.68	71.93
keystrokes (extended) +embeds	282	80.26	78.07

[3] https://bitbucket.org/vmonaco/keystroke-feature-extractor.

Using keystroke dynamics results in substantial performance gains. Already the basic feature set using 52 letter duration features clearly outperforms the stylistic features, reaching an accuracy of 71% on the test data. Adding keystroke transition durations further boost performance to 77%. These are remarkable results in light of the low baseline given by the rather large number of candidate authors. In fact, as the number of authors increases, authorship attribution becomes increasingly more difficult (Luyckx & Daelemans 2008).

4 Discussion

Our results show that the biometric keystroke features are more predictive of authorship than the stylometric features. This confirms earlier findings (Stewart et al. 2011), however, they used a simpler setup (binary classification). However, it is not straightforward to combine these two sources of information. Adding plain n-gram features (character or word n-grams) results in performance drops. In that case we add a high-dimensional sparse feature space to the dense duration feature, most probably these large amount of features swamp the feature space. In contrast, if we use word embeddings, we model the user's text as average point in a high-dimensional space and effectively add a dense low-dimensional vector to the keystroke dynamics data. This gives at times slight improvements, albeit not significant on our relatively small test and development set.

To examine which kind of features are highly predictive, we train a logistic regression model on our best configuration (extended keystrokes and embeddings) and examine the most predictive features. We see that mostly duration features of non-letter symbols are among the most predictive features, in particular punctuation symbols and spaces. This is intuitively pleasing, as users exhibit different behavior at word and sentence boundaries (Wengelin 2006).

5 Related Work

Authorship attribution has a long tradition dating back to early works in the 19th century. The most influential work on authorship attribution goes back to Mosteller & Wallace (1964), who construe it as a text classification problem (Nerbonne 2007). For a long time statistical approaches to authorship attribution focused on distributions of *function words*, high-frequency words that are presumably not consciously manipulated by the author (Nerbonne 2007; Pennebaker 2011). For example, in the well-known Federalist papers, *enough* and *while* were used exclusively by Hamilton, while *whilst* was prototypical for Madison. An early study using neural networks to infer the author of the disputed documents of the Federalist papers used 11 function words as predictive features (Tweedie, Singh & Holmes 1996). Recent work also includes authorship studies on microblog texts (Rappoport & Koppel 2013). An excellent recent summary is Stamatatos (2009).

Keystroke logging has developed into a promising tool for research into writing (Wengelin 2006; van Waes, Leijten & van Weijen 2009; Baaijen, Galbraith & de Glopper 2012), as time measurements can give insights into cognitive processes involved in writing (Nottbusch, Weingarten & Sahel 2007) or translation studies. In fact, most prior work that uses keystroke logs focuses on experimental research. For example, Hanoulle, Hoste & Remael (2015) study whether a bilingual glossary reduces the working time of professional translators. They consider pause durations before terms extracted from keystroke logs and find that a bilingual glossary in the translation process of documentaries reduces the translators' workload. Other translation research has combined eye-tracking data with keystroke logs to study the translation process (Carl et al. 2016). An analysis of users' typing behavior was studied by Baba & Suzuki (2012). They collect keystroke logs of online users describing images to measure spelling difficulty. They analyzed corrected and uncorrected spelling mistakes in Japanese and English and found that spelling errors related to phonetic problems remain mostly unnoticed.

It has been shown that pauses reflect the planning of the unit of text itself (Baaijen, Galbraith & de Glopper 2012) and that they correlate with clause and sentence boundaries (Spelman Miller & Sullivan 2006). Goodkind & Rosenberg (2015) investigate the relationship between pre-word pauses and multi-word expressions. They found that within MWE pauses vary depending on the cognitive task. Taking writing research as a starting point, a recent study postulated that keystrokes contain fine-grained information that aids the identification of syntactic chunks (Plank 2016). They integrated automatically derived labels from keystroke logs as auxiliary task in a multi-task setup (Plank, Søgaard & Goldberg 2016) with promising results. Instead, this paper focuses on the idiosyncrasy of keystroke patterns. Our results show that keystroke biometrics are far superior to that of a stylometry-based approach to authorship attribution. At the same time it is challenging to combine the two sources of information. This confirms earlier findings by the most related study (Stewart et al. 2011). They combine keystroke log features and linguistic stylometry features for user verification using a *k*-nearest neighbor approach. Their study differs from ours in two aspects. First, they use different stylometric features, i.e., the number of a specific set of characters, number of words of a certain length, average word length and number of punctuation symbols, see the full list in the appendix of their paper. Second, they target user authentication, thus their setup is a binary classification task (authenticated vs. not-authenticated), while we here focus on a multi-class classification setup (who wrote the piece of text out of all possible authors).

6 Conclusions

We have shown that keystroke dynamics contain highly indicative information to predict the authorship of a text. We compared keystroke dynamics to more traditional authorship attribution features and found keystroke biometrics to be superior. In particular, duration features of punctuation and spaces are highly predictive of authorship. However, combining keystrokes and linguistic features, two very different feature spaces, proves difficult. Some promising initial results are obtained by using word embeddings, however, further investigations are needed to test the robustness of this direction.

References

Baaijen, Veerle M., David Galbraith & Kees de Glopper. 2012. Keystroke analysis: reflections on procedures and measures. *Written Communication.*

Baba, Yukino & Hisami Suzuki. 2012. How are spelling errors generated and corrected?: a study of corrected and uncorrected spelling errors using keystroke logs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers*, vol. 2.

Carl, Michael, Isabel Lacruz, Masaru Yamada & Akiko Aizawa. 2016. Measuring the translation process. In *The 22nd Annual Meeting of the Association for Natural Language Processing.*

Collobert, Ronan, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa & Michael Collins. 2011. Natural language processing (almost) from scratch.

Goodkind, Adam & Andrew Rosenberg. 2015. Muddying the multiword expression waters: how cognitive demand affects multiword expression production. In *Proceedings of MWE 2015.*

Hanoulle, Sabien, Véronique Hoste & Aline Remael. 2015. The translation of documentaries: can domain-specific, bilingual glossaries reduce the translators' workload? an experiment involving professional translators. *New Voices in Translation Studies* (13).

Locklear, Hilbert, Sathya Govindarajan, Zdenka Sitova, Adam Goodkind, David Guy Brizan, Andrew Rosenberg, Vir V. Phoha, Paolo Gasti & Kiran S. Balagani. 2014. Continuous authentication with cognition-centric text production and revision features. In *Biometrics (IJCB), 2014 IEEE International Joint Conference.*

Luyckx, Kim & Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1, 513–520. Association for Computational Linguistics.

Monaco, John V., John C. Stewart, Sung-Hyuk Cha & Charles C. Tappert. 2013. Behavioral biometric verification of student identity in online course assessment and authentication of authors in literary works. In *Biometrics: theory, applications and systems (BTAS), 2013 IEEE Sixth International Conference on.*

Mosteller, Frederick & David Wallace. 1964. Inference and disputed authorship: the Federalist.

Nerbonne, John. 2007. The exact analysis of text. *Foreword to the 3rd edition of Frederick Mosteller and David Wallace Inference and Disputed Authorship: The Federalist Papers CSLI: Stanford.*

Nottbusch, Guido, Rüdiger Weingarten & Said Sahel. 2007. From written word to written sentence production. *Writing and cognition: Research and applications.* Mark Torrance, Luuk van Waes & David W. Galbraith (eds.). 31–54.

Pennebaker, James W. 2011. Using computer analyses to identify language style and aggressive intent: the secret life of function words. *Dynamics of Asymmetric Conflict* 4(2). 92–102.

Plank, Barbara. 2016. Keystroke dynamics as signal for shallow syntactic parsing. In *The 26th International Conference on Computational Linguistics (COLING).*

Plank, Barbara, Anders Søgaard & Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *ACL.*

Rappoport, Roy Schwartz Oren Tsur Ari & Moshe Koppel. 2013. Authorship attribution of micro-messages. In *EMNLP.*

Al-Rfou, Rami, Bryan Perozzi & Steven Skiena. 2013. Polyglot: distributed word representations for multilingual NLP. In *CoNLL.*

Spelman Miller, Kristyan & Kirk P. H. Sullivan. 2006. *Keystroke logging: an introduction.* Elsevier.

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

Stewart, John C., John V. Monaco, Sung-Hyuk Cha & Charles C. Tappert. 2011. An investigation of keystroke and stylometry traits for authenticating online test takers. In *Biometrics (IJCB), 2011 International Joint Conference.*

Sullivan, Kirk P. H., Eva Lindgren, et al. 2006. *Computer keystroke logging and writing: methods and applications.* Elsevier.

Tappert, Charles C., Sung-Hyuk Cha, Mary Villani & Robert S. Zack. 2010. A keystroke biometric system for long-text input. *International Journal of Information Security and Privacy (IJISP)* 4. 32–60.

Tweedie, Fiona J., Sameer Singh & David I. Holmes. 1996. Neural network applications in stylometry: the Federalist papers. *Computers and the Humanities* 30(1). 1–10.

van Waes, Luuk, Mariëlle Leijten & Daphne van Weijen. 2009. Keystroke logging in writing research: observing writing processes with inputlog. *GFL-German as a foreign language* 2(3). 41–64.

Wengelin, Åsa. 2006. Examining pauses in writing: theory, methods and empirical data. *Computer key-stroke logging and writing: methods and applications.* (Studies in Writing) 18. 107–130.