Conversationally-inspired Stylometric Features for Authorship Attribution in Instant Messaging

Marco Cristani
University of Verona (IT)
Istituto Italiano di Tecnologia, Genova (IT)
marco.cristani@univr.it

Giorgio Roffo
Istituto Italiano di Tecnologia, Genova (IT)
giorgio.roffo@iit.it

Cristina Segalin
Istituto Italiano di Tecnologia, Genova (IT)
criistina.segalin@gmail.com

Loris Bazzani
Istituto Italiano di Tecnologia, Genova (IT)
loris.bazzani@iit.it

Alessandro Vinciarelli
University of Glasgow (UK)
Idiap Research Institute (CH)
vincia@dcs.gla.ac.uk

Vittorio Murino
Istituto Italiano di Tecnologia, Genova (IT)
vittorio.murino@iit.it

ABSTRACT

Authorship attribution (AA) aims at recognizing automatically the author of a given text sample. Traditionally applied to literary texts, AA faces now the new challenge of recognizing the identity of people involved in chat conversations. These share many aspects with spoken conversations, but AA approaches did not take it into account so far. Hence, this paper tries to fill the gap and proposes two novelties that improve the effectiveness of traditional AA approaches for this type of data: the first is to adopt features inspired by Conversation Analysis (in particular for turn-taking), the second is to extract the features from individual turns rather than from entire conversations. The experiments have been performed over a corpus of dyadic chat conversations (77 individuals in total). The performance in identifying the persons involved in each exchange, measured in terms of area under the Cumulative Match Characteristic curve, is 89.5%.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing]. **General Terms:** Experimentation. **Keywords:** Authorship Attribution, Biometry, Instant Messaging, Chat, Social Media, Stylometry.

1. INTRODUCTION

Authorship Attribution (AA) is the domain aimed at automatically recognizing the author of a given text sample. Common techniques use *stylometric* cues that can be split into five major groups: *lexical, syntactic, structural, content-specific and idiosyncratic* [1]. Earlier approaches for automatic AA focused on printed material, typically books, and mainly exploited lexical (e.g., the frequency of characters and words) and syntactic features (e.g., punctuation, articles, propositions and other functional words) [7]. The diffusion of Internet has shifted the AA attention towards online texts (web pages, blogs, etc.) electronic messages (e-mails, tweets, posts, etc.), and other types of written information that are far shorter than an average book, way more informal and much richer in terms of expressive elements like colors, layout structures, fonts, graphics, emoticons, etc. Efforts to take into account such aspects at the level of both structure and syntax were reported in [5]. In addition, content-specific and idiosyncratic cues (e.g., topic models and grammar checking tools) were introduced to unveil deliberate stylistic choices [3].

Nowadays, one of the most important AA challenges is the identification of people involved in chat (or chat-like) conversations. The task has become important after that social media have penetrated the everyday life of many people and have offered the possibility of interacting with persons that hide their identity behind nick-names or potentially fake profiles. So far, standard stylometric features have been employed to categorize the content of a chat [11] or the behavior of the participants [15], but attempts of identifying chat participants are still few and early. Furthermore, the similarity between spoken conversations and chat interactions has been neglected while being a key difference between chat data and any other type of written information.

Hence, this paper proposes a set of novel stylometric features that take into account the conversational nature of chat interactions. Some of them fit in the taxonomy proposed at the beginning of this section, but others require to define the new group of *conversational* features. The reason is that they are based on *turn-management*, probably the most salient aspect of spoken conversations that applies to chat interactions as well. In conversations, turns are intervals of time during which only one person talks. In chat interactions, a turn is a block of text written by one participant during an interval of time in which none of the other participants writes anything. Like in the case of automatic analysis of spoken conversations, AA features are extracted from individual turns and not from the entire conversation.

The experiments are performed over a corpus of dyadic chat conversations that involve 77 subjects. The average number of available words per subject is 615 and the AA performance, measured with the area under the Cumulative Match Characteristic curve, is 89.5%.

The rest of the paper is organized as follows: Section 2

MM'12, October 29–November 2, 2012, Nara, Japan.

sketches the literature on application of AA to chat data; Section 3 presents the new features proposed in this work; Section 4 reports on experiments and results and the final Section 5 draws conclusions and outlines future perspectives.

2. RELATED WORK

The application of AA to chat conversations is recent (see [12] for a survey). Typically, state-of-the-art approaches extract stylometric features from the data and use discriminative classifiers to identify the author (each author corresponds to a class). Table 1 is a synopsis of the features applied so far in the literature. The extraction process is always applied to the entire conversation and individual turns, while being the basic blocks of the conversation, are never used as analysis unit. In [14], lexical, syntactic, structural and content-specific features are fed to SVM, Neural Networks and decision trees on 20 subjects. The work in [2] adds idiosyncratic features to the previous ones and applies a PCA-like projection onto a low-dimensional, but highly discriminant space to identify 100 potential authors of e-mails, Instant Messages, feedback comments and program code (thousands of words per identity). The problem of the size (i.e., the number of conversations per author) of the training set has been studied in [8], with different classifiers and 20 subjects. Special characters (e.g., emoticons and abbreviations) have been exploited in [11] with decision trees, K-nearest-neighbor and naive Bayes classifiers for discriminating 4 subjects.

The main limitation of the works above is that they do not process chat exchanges as conversations, but as normal texts. This work tries to overcome such limitation and introduces a new class of features that account for the presence of turns (see below) in chat conversations. Furthermore, the proposed approach does not apply the feature extraction process to the entire conversation (like in all works above), but to individual turns.

3. FEATURE EXTRACTION

The data set includes $N = 77$ subjects, each involved in a dyadic chat conversation with an interlocutor. The conversations can be modeled as sequences of *turns*, where "*turn*" means a stream of symbols and words (possibly including "return" characters) typed consecutively by one subject without being interrupted by the interlocutor. The feature extraction process is applied to T consecutive turns that a subject produces during the conversation. Because of privacy and ethical issues, features that do not involve the content of the conversation can be used, namely number of words, characters, punctuation marks and emoticons.

In standard AA approaches, the features are extracted from a chat conversation as a whole. In our case, we extract the features from each turn individually and then we estimate their statistics (mean value, histograms, etc.). In particular, we opted for exponential histograms where the size of the bins changes according to the value of the feature (bins are smaller for small values and larger for higher values). The reason is that the turns are short and small values tend to be more represented. This choice is shown to improve the performance (see below).

The introduction of turns as a basic analysis unit allows one to introduce features that explicitly take into account the conversational nature of the data and mirror behavioral measurements typically applied in automatic understanding of social interactions (see [13] for an extensive survey):

- **Turn duration**: the time spent to complete a turn (in hundredth of seconds); this feature accounts for the rhythm of the conversation with faster exchanges typically corresponding to higher engagement.
- **Writing speed** (two features): number of typed characters -and words- per second (typing rate); these two features indicate whether the duration of a turn is simply due to the amount of information typed (higher typing rates) or to cognitive load (low typing rate), i.e. to the need of thinking about what to write
- **Number of "return" characters**: since these latter tend to provide interlocutors with an opportunity to start a new turn, high values of this feature are likely to measure the tendency to hold the floor and prevent others from "speaking" (an indirect measure of dominance).
- **Mimicry**: ratio between number of words in current turn and number of words in previous turn; this feature models the tendency of a subject to follow the conversation style of the interlocutor (at least for what concerns the length of the turns). The mimicry accounts for the social attitude of the subjects.

We call these features *conversational* features.

No.	Feature	Range
1	# words	[0,260]
2	# emoticons	[0,40]
3	# emoticons per word	[0,1]
4	# emoticons per characters	[0,0.5]
5	# exclamation marks	[0,12]
6	# question marks	[0,406]
7	# characters	[0,1318]
8	average word length	[0,20]
9	# three points	[0,34]
10	# uppercase letters	[0,94]
11	# uppercase letters/#words	[0,290]
12	**turn duration**	[0,1800(sec.)]
13	**# return chars**	[1,20]
14	**# chars per second**	[0,20(ch./sec.)]
15	**# words per second**	[0,260]
16	**mimicry degree**	[0,1115]

Table 2: Stylometric features used in the experiments. The symbol "#" stands for "number of". In bold, the conversational features.

Table 2 provides basic facts about the features used in the experiments. In the case of 1-13 and 16 the features correspond to the exponential histograms (32 bins) collected from the T turns. In the case of 14 and 15, the features correspond to the average estimated over the T turns. This architectural choice maximized the AA accuracy.

4. EXPERIMENTS

The experiments have been performed over a corpus of dyadic chat conversations collected with Skype (the language is Italian). The conversations are spontaneous, i.e.

Group	Description	Examples	References
Lexical	Word level	Total number of words (=M), # short words/M, # chars in words/C, # different words, chars per word, freq. of stop words	[2, 8, 11, 12, 14]
	Character level	Total number of characters (chars) (=C), # uppercase chars/C, # lowercase chars/C, # digit chars/C, freq. of letters, freq. of special chars	[2, 11, 12, 14]
	Character|Digit n-grams	Count of letter|digit n-gram (a, at, ath, 1 , 12 , 123)	[2, 12, 14]
	Word-length distribution	Histograms, average word length	[2, 8, 11, 12, 14]
	Vocabulary richness	Hapax legomena, dislegomena	[2, 8, 12, 14]
Syntactic	Function words	Frequency of function words (of, for, to)	[2, 8, 11, 12, 14]
	Punctuation	Occurrence of punctuation marks (!, ?, :), multiple !|?	[2, 8, 11, 12, 14]
	Emoticons|Acronym	:-), L8R, Msg, :(, LOL	[11, 12]
Structural	Message level	Has greetings, farewell, signature	[2, 8, 11, 12, 14]
Content-specific	Word n-grams	Bags of word, agreement (ok, yeah, wow), discourse markers|onomatopee (ohh), # stop words, # abbreviations , gender|age-based words, slang words	[2, 8, 11, 12, 14]
Idiosyncratic	Misspelled word	Belveier instead of believer	[2, 8, 11, 12]

Table 1: Synopsis of the state-of-the-art features for AA on chats. "#" stands for "number of".

they have been held by the subjects in their real life and not for the purpose of data collection. This ensures that the behavior of the subjects is natural and no attempt was made to modify the style in any sense. The number of turns per subject ranges between 60 and 100. Hence, the experiments are performed over 60 turns of each person. In this way, any bias due to differences in the amount of available material should be avoided. When possible, we selected different turns (maintaining their chronological order) in order to generate different AA trials. The average number of words per subject is 615. The 60 turns of each subject are split into *probe* and *gallery* set, each including 30 samples.

The first part of the experiments aims at assessing each feature independently as a simple ID signature. A particular feature of a single subject is extracted from the probe set, and matched against the corresponding gallery features of all subjects, employing an appropriate metric (Bhattacharya distance for histograms [6] and Euclidean distance for mean values). This happens for all the probe subjects, resulting in a $N \times N$ distance matrix (N is the total number of subjects). Ranking in ascending order the N distances for each probe element allows one to compute the *Cumulative Match Characteristic* (CMC) curve, i.e., the probability of finding the correct match in the top n positions of the ranking (with n ranging between 1 and N). The CMC curve is an common performance measure for AA approaches [4]. In particular, the value of the CMC curve at position 1 is the probability that the probe ID signature of a subject is closer to the gallery ID signature of the same subject than to any other gallery ID signature; the value of the CMC curve at position n is the probability of finding the correct match in the first n ranked positions.

Given the CMC curve for each feature (obtained by averaging on all the available trials), the normalized Area Under Curve (nAUC) is calculated as a measure of accuracy. Figure 1 shows that the individual performance of each feature is low (less than 10% at rank 1 of the CMC curve). In addition, the first conversational feature ranks seventh in terms of nAUC, while the others rank 10^{th}, 14^{th}, 15^{th} and 16^{th}, respectively.

The experiments above serve as basis for the Forward Feature Selection (FFS) strategy [10]. At the first iteration the FFS retains the feature with the highest nAUC, at the second one it selects the feature that, in combination with

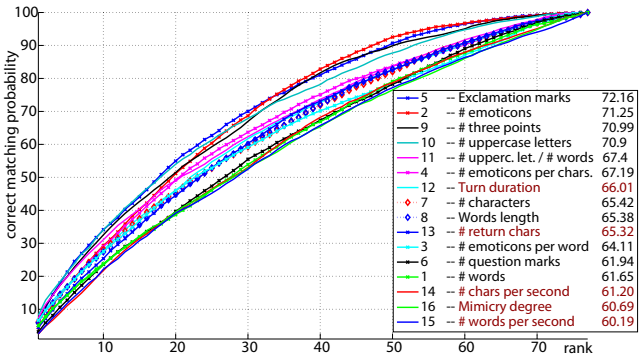

Figure 1: CMCs of the proposed features. The numbers on the right indicate the nAUC. Conversational features are in bold (best viewed in colors).

the previous one, gives the highest nAUC, and so on until all features have been processed. In our experiments, combining features means to average their related distance matrices, forming a composite one. The pool of selected features is the one which gives the highest nAUC. Since FFS is a greedy strategy, different runs (50) of the feature selection are used, selecting a partially different pool of 30 turns each time for building the probe set. In this way, 50 different ranked subsets of features are obtained. For distilling a single subset, the Kuncheva stability index [9] is adopted, which essentially keeps the most informative features (with high ranking in the FFS) that occurred most times.

The FFS process results into 12 features, ranked according to their contribution to the overall CMC curve. The set includes features 5, 2, 9, 10, **12 (turn duration)**, **13 (# "return" characters)**, 8, **14 (chars per second)**, 6, 7, **16 (mimicry degree)**, **15 (words per second)**. In bold, we report the conversational features that appear to rank higher than when used individually. This suggests that, even if their individual nAUC was relatively low, they encode information complementary with respect to the traditional AA features.

The final CMC curve, obtained using the pool of selected features, is reported in Figure 2, curve (a). In this case, the rank 1 accuracy is 29.2%. As comparison, other CMC curves are reported, considering (b) the whole pool of fea-

# Turns	5	10	15	20	25	30
nAUC	68.6	76.6	80.6	85.0	88.4	89.5
rank1 acc.	7.1	14.0	15.1	21.9	30.6	29.2

Table 3: Relationship between performance and number of turns used to extract the ID signatures.

tures (without feature selection); (c) the same as (b), but adopting linear histograms instead of exponential ones; (d) the selected features with exponential histograms, without the conversational ones; (e) the conversational features alone and (f) the selected features, calculating the mean statistics over the whole 30 turns, as done usually in the literature with the stylometric features.

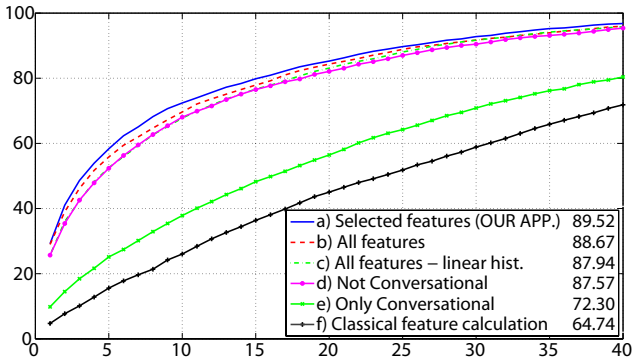

Figure 2: Comparison among different pool of features.

Several facts can be inferred: our approach has the highest nAUC; feature selection improves the performance; exponential histograms work better than linear ones; conversational features increase the matching probability of around 10% in the first 10 ranks; conversational features alone give higher performance of standard stylometric features, calculated over the whole set of turns, and not over each one of them. The last experiment shows how the AA system behaves while diminishing the number of turns employed for creating the probe and gallery signatures. The results (mediated over 50 runs) are shown in Table 3. Increasing the number of turns increases the nAUC score, even if the the increase appears to be smaller around 30 turns.

5. CONCLUSIONS

This paper proposes two main contributions to the problem of recognizing automatically the identity of chat participants while respecting their privacy. The first is the introduction of new features that account for turn-taking and mirror the features typically applied in automatic understanding of spoken conversations. The second is the use of turns as a basic analysis unit for the analysis of chat data and identification of their participants. The results are promising and show that taking into account the conversational nature of the texts typed during chat exchanges can improve the performance of AA approaches. Future work will aim not only at further exploiting such an aspect of chat conversations, but also at using more sophisticated statistical models for identity recognition.

Acknowledgments.

The work by A.Vinciarelli was supported by the the European Commission via the grant agreement 231287 (SSPNet) and by the Swiss National Science Foundation via IM2.

6. REFERENCES

[1] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2):1–29, 2008.

[2] A. Abbasi, H. Chen, and J. Nunamaker. Stylometric identification in electronic markets: Scalability and robustness. *Journal of Management Information Systems*, 25(1):49–78, 2008.

[3] S. Argamon, M. Koppel, J. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Communications of ACM*, 52(2):119–123, 2009.

[4] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to Biometrics*. Springer Verlag, 2003.

[5] O. De Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for auhtor identification forensics. *ACM SIGMOD Record*, 30(4), 2001.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley and Sons, 2001.

[7] D. I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[8] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, 2011.

[9] L. Kuncheva. A stability index for feature selection. In *IASTED International Multi-Conference Artificial Intelligence and Applications*, pages 390–395, 2007.

[10] H. Liu and H. Motoda. *Computational Methods of Feature Selection*. Chapman and Hall, 2008.

[11] A. Orebaugh and J. Allnutt. Classification of instant messaging communications for forensics analysis. *Social Networks*, pages 22–28, 2009.

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

[13] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[14] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2006.

[15] D. Zhou, L. Zhang. Can online behavior unveil deceivers? – an exploratory investigation of deception in instant messaging. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2004.