Journal of Physics: Conference Series **410** (2013) 012096 doi:10.1088/1742-6596/410/1/012096

Entropy growth in emotional online dialogues

J Sienkiewicz[1], M Skowron[2], G Paltoglou[3] and Janusz A. Hołyst[1]

[1] Faculty of Physics, Centre of Excellence for Complex Systems Research, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland

[2] Interaction Technologies Group, Austrian Research Institute for Artificial Intelligence, Freyung 6/3/1a, A-1010 Vienna, Austria

[3] School of Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, United Kingdom

E-mail: `julas@if.pw.edu.pl`

Abstract. We analyze emotionally annotated massive data from IRC (Internet Relay Chat) and model the dialogues between its participants by assuming that the driving force for the discussion is the entropy growth of emotional probability distribution.

1. Introduction

Large amounts of data offer new possibilities of examining communication between humans expressed in face-to-face encounters [1], mobile telephone calls [2], surface-mail [3] short messages [4] as well as typical Internet activities such as e-mail correspondence [5], forum postings [6] or web browsing [7]. In the most fundamental part, the communication is based on a *dialogue* - an exchange of information and ideas between two people [8]. As compared to the off-line communication, the exchange of information in the Internet is claimed to be more biased toward the emotional aspect [9]. However, the very Internet that gives the opportunity to acquire massive data, thus making it possible to perform a credible statistical analysis of common habits in communication. As the recent research shows, it is already possible to spot and model certain phenomena of the Internet discussion participants while looking just at the emotional content of their posts [10, 11, 12, 13, 14]. One of them is the collective emotional behavior [10, 13], the other is clear correlation between the length of discussion and its emotional content [10, 11]. Here we present an approach based on the observation of entropy of emotional probability distribution during the conversation that can arguably serve as an indicator of an ending discussion.

2. Data description

As a source of data for analysing online dialogues we chose the Internet Relay Chat (IRC) [15] logs. The analysis is limited only to one of the channels, namely `#ubuntu` [16] in the period 1st January 2007 - 31st December 2009. We focused on dialogues that included only two participants. The final output, after several levels of data processing consists of $N = 93329$ dialogues with the length L between $L_{min} = 11$ and $L_{max} = 339$ each. Each dialogue can be represented as a chain of messages where all odd posts are submitted by one user and all even by another one.

The emotional classifier program that was used to analyze the emotional content of the discussions is based on a machine-learning (ML) approach. The algorithm functions in two

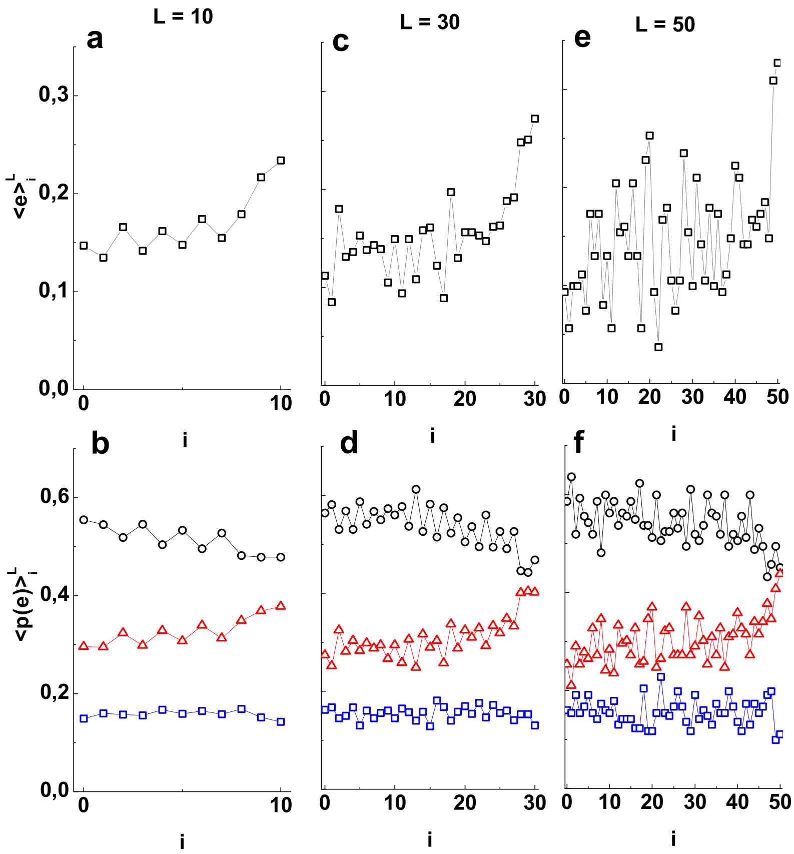

Figure 1. Average emotional value $\langle e \rangle_i^L$ (top row) and average emotional probabilities (bottom row) $\langle p(-) \rangle_i^L$ (squares), $\langle p(0) \rangle_i^L$ (circles), $\langle p(+) \rangle_i^L$ (triangles) in the i-th timestep for dialogues of specific $L = 10$ (first column), $L = 30$ (second column) and $L = 50$ (third column).

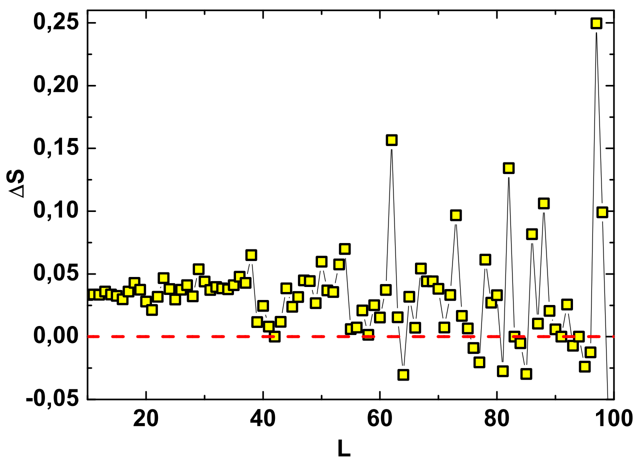

Figure 2. Difference between terminal and initial entropy value ΔS versus the dialogue length L.

phases: during the training phase, it is provided with a set of documents classified by humans for emotional content (positive, negative or objective) from which it learns the characteristics of each category. Then, during the application phase, the algorithm applies the acquired sentiment classification knowledge to new, unseen documents. In our analysis, we trained a hierarchical Language Model [17, 18] on the Blogs06 collection [19] and applied the trained model to the extracted IRC dialogues, during the application phase. The algorithm is based on a two-tier solution, according to which a post is initially classified as objective or subjective and in the latter case, it is further classified in terms of its polarity, i.e., positive or negative. Each level of classification applies a binary Language Model [20, 18]. Posts are therefore annotated with a

Table 1. Properties of the dataset: number of comments C, number of dialogues N, shortest dialogue L_{min}, longest dialogue L_{max}, average valence $\langle e\rangle$, probability of finding negative, neutral or positive emotion (respectively $p(-)$, $p(0)$ and $p(+)$).

C	N	L_{min}	L_{max}	$\langle e\rangle$	$p(-)$	$p(0)$	$p(+)$
1889120	93323	11	339	0.17	0.15	0.53	0.32

single value $e = -1, 0$ or 1 to quantify their emotional content (to be more precise - their valence [21]) as negative, neutral or positive, respectively. The accuracy of the classifier checked for 950 humanly annotated comments in IRC data is 62.49 % for subjectivity detection and 70.25 % for polarity detection.

3. Results

The obtained dialogues have been divided into groups of constant dialogue length L. For such data we follow the evolution of mean emotional value $\langle e\rangle_i^L$ and average emotional probabilities $\langle p(e)\rangle_i^L$ ($\langle e\rangle_i^L$. In both cases the $\langle ...\rangle_i^L$ symbol indicates taking all dialogues with a specific length L and averaging over all comments with number i, thus, for example, $\langle p(-)\rangle_i^L$ is the probability that at the position i in all dialogues of length L there is a negative statement. The characteristic feature observed regardless of the dialogue length is that the $\langle e\rangle_i^L$ at the end of the dialogue is higher than at the beginning (upper row in Fig. 1). In fact, there is especially a rapid growth close the very end of the dialogue, which is probably caused by participants who acknowledge others' support issuing comments like "thank you", "you were most helpful", etc.

The direct reason for such behavior is shown in the bottom row of Fig. 1, which presents the evolution of the average emotional probabilities $\langle p(-)\rangle_i^L$, $\langle p(0)\rangle_i^L$ and $\langle p(+)\rangle_i^L$. The observations can be summarized in the following way:

- the negative emotional probability $\langle p(-)\rangle_i^L$ remains almost constant,
- $\langle p(+)\rangle_i^L$ increases and $\langle p(0)\rangle_i^L$ has an opposite tendency,
- $\langle p(+)\rangle_i^L$ and $\langle p(0)\rangle_i^L$ tend to equalize in the vicinity of dialogue end.

Other manifestation of the system's features can be spotted by examining the level of the entropy S of the emotional probabilities $\langle p(e)\rangle_i^L$. Basing on entropy, it has also been shown how the coherent structures in the e-mail dialogues arise [5] or how to predict conversation patterns in face-to-face meetings [1]. It is also possible to use other information theoretic quantities as Kullback-Leiber or Jensen-Shannon divergence (see e.g., [22]), however due to an established connection to its thermodanymical equivalent as well as its clear role for poiting the direction of physical processes we use entropy (after Shannon's definition [23])

$$S_i^{sh} = -\sum_{e=-1,0,1} \langle p(e)\rangle_i^L \ln\langle p(e)\rangle_i^L. \tag{1}$$

Here, taking into account the fact that $\langle p(-)\rangle_i^L$ is constant in the course of dialogue, we paid attention only to $\langle p(+)\rangle_i^L$ and $\langle p(0)\rangle_i^L$, thus the observed entropy had a form of

$$S_i = -\left[\langle p(0)\rangle_i^L \ln\langle p(0)\rangle_i^L + \langle p(+)\rangle_i^L \ln\langle p(+)\rangle_i^L\right]. \tag{2}$$

Plotting the difference between terminal and initial entropy ΔS versus the length of the dialogue L it is possible to see that for the dialogues up to $L \approx 50$ this difference is always above zero (see Fig. 2). Over that point the system we observe high fluctuations due to small statistics of number of dialogues (dialogues with length $L > 50$ make around 1% of the total number of dialogues N) . It implies a following likely scenario for the dialogue: it evolves in the direction of growing entropy. In the beginning of the dialogue, the probabilities $\langle p(0)\rangle_i^L$ and $\langle p(+)\rangle_i^L$ are separated from each other, contributing to low value of initial entropy S_p. However, then the entropy grows, the probabilities $\langle p(0)\rangle_i^L$ and $\langle p(+)\rangle_i^L$ equalize leading to high value entropy (i.e., higher then the initial one) at the end of the dialogue.

4. Conclusions

Analysis performed on the emotionally annotated dialogues extracted from IRC data demonstrate that following such simple metrics as probability of specific emotion can be useful to predict the future evolution of the discussion. Moreover, all the analyzed dialogues share the same property, i.e., the tendency to evolve in the direction of a growing entropy. Those observations may be of help for designing the next generation of interactive software tools [24]

Acknowledgments

This work was supported by a European Union grant by the 7th Framework Programme, Theme 3: Science of complex systems for socially intelligent ICT. It is part of the CyberEmotions (Collective Emotions in Cyberspace) project (contract 231323). J.S. and J.A.H. also acknowledge support from Polish Ministry of Science Grant 1029/7.PR UE/2009/7. The work was partially supported from polish Ministry of Science and Higher Education (grant no 0490/IP3/2011/71).

References

[1] Takaguchi T, Nakamura M, Sato N, Yano K and N. Masuda 2011 *Phys. Rev.* X **1** 011008.
[2] Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész K and Barabási A-L 2007 *Proc. Net. Acad. Sci. U.S.A.* **104** 7332.
[3] Oliveira J G and Barabási A-L 2005 *Nature (London)* **437** 1251.
[4] Wu Y, Zhou Ch, Xiao J, Kurths J, and Schellnhuber H J 2010 *Proc. Nat. Acad. Sci. U.S.A.* **107** 18803.
[5] Eckmann J-P, Moses E and Sergi D 2004 *Proc. Nat. Acad. Sci. U.S.A.* **101** 14333.
[6] Kujawski B, Hołyst J A and Rodgers G J 2007 *Phys. Rev.* E **76** 036103.
[7] Chmiel A, Kowalska K and Hołyst J A 2009 *Phys. Rev.* E **80** 066122.
[8] Sacks H, Schegloff E A and Jefferson G 1974 *Language* **50** 696.
[9] Sobkowicz P and Sobkowicz A 2010 *Eur. Phys. J.* B **73**, 633.
[10] Chmiel A, Sienkiewicz J, Thelwall M, Paltoglou G, Buckley K, Kappas A and Hołyst J A 2011 *PLoS ONE* **6** e22207.
[11] Chmiel A, Sobkowicz P, Sienkiewicz J, Paltoglou G, Buckley K, Thelwall M and Hołyst J A 2011 *Physica* A **390** 2936.
[12] Weroński P, Sienkiewicz J, Paltoglou G, Buckley K, Thelwall M and Hołyst J A 2012 *Acta Phys. Pol.* A **121** B-128.
[13] Garas A, Garcia D, Skowron M and Schweitzer F 2012 *Sci. Rep.* **2** 402.
[14] Garcia D, Garas A and Schweitzer F 2012 *EPJ Data Science* **1** 3.
[15] http://en.wikipedia.org/wiki/Internet_Relay_Chat.
[16] https://help.ubuntu.com/community/InternetRelayChat.
[17] Sebastiani F 2002 *ACM Computing Surveys* **34**.
[18] Mitrović M, Paltoglou G and Tadić B 2010 *Eur. Phys. J.* B **77**, 597.
[19] Ounis I, Macdonald C, Soboroff I 2008 in *Proceedings of the Second International Conference on Weblogs and Social Media.*
[20] Peng F, Schuurmans D, and Wang S 2003 in *NAACL '03* 110.
[21] Feldman L A 1995 *Journal of Personality and Social Psychology* **69** 153.
[22] Masucci A P, Kalampokis A, Eguíluz V M and Hernández-García E 2011 *Phys. Rev. E* **83**, 026103.
[23] Shannon C E 1948 *Bell Syst. Tech. J.* **27** 379.
[24] Skowron M, Rank S, Theunis M and Sienkiewicz J 2011 *LNCS* **6974** 337.