Toward the Discovery of Chances in the Analysis of Natural Language

Calkin A.S. Montero and Kenji Araki

Graduate School of Information Science and Technology, Hokkaidō University,
Kita 14-jo Nishi 9-chome, Kita-ku, Sapporo, 060-0814 Japan
{calkin,araki}@media.eng.hokudai.ac.jp

Abstract. Chance discovery is being applied to a wide range of branches of science fundamentally when there are events that are difficult to find or predict. In this paper we present a summary of the application of chance discovery techniques to the processing of natural language, focusing on the analysis of human-computer chat.

1 Introduction

Chance discovery, as a field of investigation itself, is devoted to analyze the occurrence of events that might have important impact in their happening area. Despite of being a new field of research, the ubiquity of chance discovery is continuously growing. In all of the applications, the word "chance" has been considered as an event that might have an important impact in its happening area. Hence, chance discovery is dedicated to analyse the occurrence of such events that might represent rare opportunities - or risks - whose significance has been unnoticed, and which may lead to an unexpected benefit - or catastrophe.

Chances have been studied in areas of application such as: earthquake prediction, genomes triggering diseases, complex adaptive systems, discourse analysis, and so forth (for more details see [1]). In the same tradition, this paper presents the analysis of chances within human-computer chat[1]. As to define "chance" in this research, let us refer to the analysis of human-human chat.

When analyzing human-human chat a chance can be regarded as the moment in the dialogue when a topic-shift occurs. A topic-shift can be considered as the transition from one discussed subject to another during the dialogue. Previous research [2] has shown a peculiar behavior of critical self-organization of the human-human chat. This behavior can be considered to be triggered by the positive feedback effect that exists from the interlocutors during the chat as a result of the topic-shifting dynamics. We argue that the importance of understanding the critical dynamics of human-human chat can become relevant when modeling human-like computer chat.

This paper presents an approach to analyse natural language through the application of chance discovery techniques, the KeyGraph and data crystallization, as to observe the dynamics of human-computer chat when chances are missing.

[1] Chat: friendly, informal conversation. Cambridge Advanced Learner's Dictionary

2 Natural Language and Chance Discovery

While researching on the structure of human-human chat, a peculiar general behavior of self-organization has been observed [2]. This behavior has been explained using an analogy with the self-organized criticality of the sandpile model described by Bak in [3]. In order to explain the observed behavior in the flow of a conversation, let us consider our system to be a human-human chat section. By analogy with the sandpile model we could state that in the beginning, the chat could be considered flat (greetings from each interlocutors). As the chat evolves toward one direction - some specific topic - certain utterances may cause the course of the chat to slightly chance the direction of the topic, although the initial topic still remains the same. Eventually, there is nothing else to utter about the initial topic, i.e., because the interlocutors have already agreed in their ideas or the topic has lost its novelty. At this point a single utterance given by any of the interlocutors can cause the whole direction of the chat to change toward a new topic - i.e., topic-shifting, representing the span of the sandpile - and the whole process starts again with a new topic. Due to the dynamics of this cyclic behavior the chat has been considered to show self-organization, i.e., Critically Self-Organized Chat. As a next step, in order to observe and understand the dynamics of the human-computer chat the tools described below were used.

2.1 The KeyGraph

In order to understand the behavior of possible chance events, visual computer aid is needed. In this regard the KeyGraph has raised as a data-mining tool for extracting patterns of the appearance of chance events. The KeyGraph tool was originally used for indexing documents [4]. As a data-mining algorithm, the KeyGraph identifies relationships between terms in a document particularly focusing on co-occurrence relationships of both high-probability and low-probability events. The KeyGraph has been applied to a variety of topics including the analysis of documents of speech, discovering deep building blocks for genetic algorithms, analysis of point of sale (POS) data, earthquake prediction , and so forth. In order to observe the behavior and characteristics of human-computer chat, the interrelationship of utterances of a chat section[2] was analyzed following the procedure: a) each utterance was considered one sentence, b) each sentence was segmented into words, c) high frequency words were eliminated, i.e., I, you, is, and the like, d) sentences co-occurrence relationship was determined as:

D= w_1:: S_1, S_2, S_4 .../ w_2:: S_9, S_{25} .../ w_3:: S_1, S_3, S_{10} .../ ... /
w_n:: S_{24}, S_{25}, ... S_m

where: w_k $(k = 1, 2, 3, ..., n)$, represents a word in a sentence.
S_l $(l = 1, 2, 3, ..., m)$, represents a sentence.

The obtained D document contains the co-occurrence relationship of the utterances during the analyzed chat section. This D document is then process with another chance discovery technique, data crystallization.

[2] Using recorded chat section between a user and ALICE chatbot

2.2 Data Crystallization

Data crystallization, as stated in [5], is dedicated to experts working in real domains where discoveries of unobservable events are desired. Data crystallization as an extension of chance discovery tries to reveal events that are important but are not observed since they are not included in the analyzed data.

Data crystallization has been proposed to be applied for identifying missing links, i.e., hidden leaders among organizations. Fundamentally, dummy events are inserted to the given data, named as X_Y, where X represents the level in the hierarchical structure of the organization and Y the line in the original data where the dummy item was inserted. After inserting the dummy items into the given data, the KeyGraph is applied and all the dummy nodes that did not appear linking clusters in the resulting graph are eliminated from the data, and then the cycle is iterated, i.e., insertion of dummy events, to higher levels. Unobservable events and their relations with the observable events given in the original data are to be visualized by the application of KeyGraph.

We inserted dummy items to the co-occurrence relationship document, D document, showed above, as follows:

D'= w_1:: S_1, S_2, S_4 ...1_1 / w_2:: S_9, S_{25} ...1_2 / w_3:: S_1, S_3, S_{10} ...1_3 / ... / w_n:: S_{24}, S_{25}, ... S_m 1_n

where: 1_o ($o = 1, 2, 3, ..., n$), represents each dummy item inserted in each line of the original data.

After feeding the KeyGraph with this D' document, all the dummy items that did not appear linking clusters as bridges in the outputted graph are deleted from the document. Here new dummy items with higher hierarchy are inserted in the data, iterating the cycle.

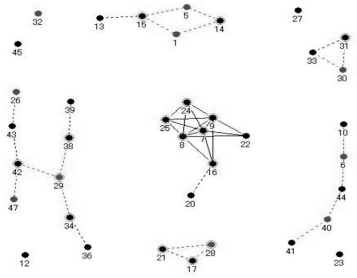
3 Experimental Results

During the experiment, a chat section between the chatbot and a native English speaker was analyzed in order to find co-occurrence between the user's utterance and the chatbot replies. There were 48 turns of the interlocutors (user - chatbot). The document of utterance co-occurrence was obtained as described in Sect. 2.1 and was examined by the KeyGraph. Fig. 1 shows the graphical view of the chat section.

In this figure, the clusters represent the relationship between the interlocutors utterances, and the links between clusters represent the transition between one topic and the other. It can be observed that the main clusters are not interconnected, meaning the chatbot in many cases could not keep a smooth and natural flow of the chat, giving as a result vagueness during the dialogue.

As a continuation of the experiment, the same chat section was analyzed following the procedure of data crystallization described in Sect. 2.2. The graphical output can be observed in Fig. 2.

This figure shows how the two main clusters appear interconnected by the dummy item 1_3. This can be regarded as the need of a *missing* critical utterance,



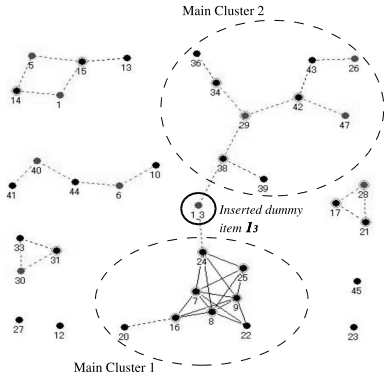


Fig. 1. Non-Critical Chat Graphical View **Fig. 2.** Observation of a Missing Chance

i.e. chance, after the utterance 24 in order to shift smoothly the topic toward a new direction. In doing so, a well interconnected graph could be obtained, which represent a smooth flow of the dynamics of the chat, as it is the case of human-human chat. With this preliminary experiment the dynamics of the human-computer chat could be appreciated, being the lack of criticality of it revealed. As to observe the critical behavior - or lack of it - of human-computer chat this approach showed to be valid.

4 Conclusion

In this paper the application of chance discovery techniques to the analysis of natural language has been presented. The results given by the KeyGraph showed a clear difference in the structure of the chat flow in concordance of the presence or absence of a critical utterance. Future work will be oriented toward the enhancing of the modeling of human-computer chat.

References

1. Ohsawa Y, McBurney P: Chance Discovery. Springer, Berlin Heidelberg New York. (2003)
2. Montero C.A.S., Araki K.: Human Chat and Self-Organized Criticality: A Chance Discovery Application. New Mathematics and Natural Computation, Vol. 1, No. 3 pp 407-420 (2005)
3. Bak P: How Nature Works, Oxford University Press. (1997)
4. Ohsawa Y, Benson NE, Yachida M: KeyGraph: Automatic Indexing by Co-occurrence Graph Based on Building Construction Metaphor. Proceedings of Advanced Digital Library Conference, pp. 12-18. (1998)
5. Ohsawa Y: Data Crystallization: Chance Discovery Extended for Dealing with Unobservable Events New Mathematics and Natural Computation, Vol. 1, No. 3 pp 373-392 (2005)