Interaction Pattern and Motif Mining Method for Doctor-Patient Multi-Modal Dialog Analysis

Kenji Mase[1] Yuichi Sawamoto[1]* Yuichi Koyama[1]
Tomio Suzuki[2] Kimiko Katsuyama[3]
[1]Graduate School of Information Science, Nagoya University, Nagoya 464-8601 Japan
[2]Nagoya University Hospital, Nagoya 466-8560 Japan
[3]Graduate School of Nursing, Osaka Prefecture University, Osaka 583-8555 Japan
{koyama, mase}@nagoya-u.jp, +81-52-789-5898

ABSTRACT

We propose a bottom-up analysis method of multi-modal dialogue interaction with a pattern and motif mining method to summarize such interviews as between doctors and patients for medical diagnosis. Our aim is to generate a hierarchical model of the interviewing behavior of such kinds as interaction corpora, consisting of primitive, pattern, motif, and pattern clusters from the given dialogue session data. We exploit a Jensen-Shannon Divergence measure to extract important patterns and motifs. Medical interview is chosen as an important application of such analysis because a doctor's multi-modal interviewing technique is essential to establish a reliable relationship and to conclude with a successful diagnosis.

An interaction corpus of example simulated medical interviews is constructed by the proposed method. The interviews are captured by a video camera and microphones. Based on the constructed indices in terms of given pattern notations and clusters, the interviews were summarized. Performance evaluation of the indices by a medical doctor was performed to confirm their plausibility and summary descriptions of the results.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine System—*Human Information Processing*

General Terms

Human Factors, Algorithms, Experimentation

Keywords

Medical Dialogue, Multi-modal Interaction, Data Mining, Gaze, Gesture, Jensen-Shannon Divergence

*Now working for KDDI Corporation, JAPAN

ICMI-MLMI'09, Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing. November 6, 2009, Boston, MA USA

Figure 1: Simulated Medical Interview Lesson between a medical student and a simulated patient

1. INTRODUCTION

We propose a corpus development method for multi-modal interaction such as general dialogs, medical interviews, and office meetings where verbal and non-verbal interactions occur for mutual understanding or cooperative creative work. Such an interaction corpus is constructed by annotating the input signals of people's behavior along with the environmental situation and the machine's responses. A valuable corpus should contain not only such primitive indices as speech, gaze, and location events but also such higher level abstract indices as eye-contact, key action, dominant speaking events, etc. Such a higher level index can convey human values and the meanings of actions.

We propose a bottom-up, hierarchical and un-supervised method of indice generation for a corpus in this paper. It uses a large-scale data set to include various possible pattern events for consideration and to reduce them to representative, higher level notations. We newly introduce *Motif* and its crucial measure as sequential pattern indices and exploit the measure to obtain a distance measure to construct *Pattern Clusters*. The Jensen-Shannon Divergence is exploited to give these measures of the important patterns and motifs from the given data set.

Validity evaluation of the index design is an important issue, especially for a bottom-up approach design. In this paper, we apply this method to medical interview sessions between doctors and patients and evaluate its validity (see Figure 1). Medical interview is a hot topic in medical education, and the interview technique analysis is being performed in various ways. We recorded many simulated interview sessions with medical students and simulated patients for analysis by means of conventional sensors; video cam-

eras and microphones. The indices are developed by the proposed method, and the summaries of the interview interactions are evaluated by a physician to confirm the method's validity and the adequacy of the obtained summary.

2. RELATED WORKS

There are many research works on human behavior recognition and understanding. Theory and technology are needed for modeling the structure of human-machine and human-human interaction for future ubiquitous information environments and the human-robot symbiotic world [1].

For example, Wren et al. [2] used more than 200 motion sensors on the ceiling with a Markov process to find such events in offices as special visits by executives. Kanda et al. [3] extracted typical behavioral patterns of museum visitors using DP matching and clustering. Both works modeled individual behaviors based on the sensing data of real activities. Otsuka et al. [4] proposed a probabilistic framework for inferring the structure of conversations in face-to-face multiparty communication based on gaze patterns, head directions, and the presence/absence of utterances. They defined the typical types of conversation structure in a top-down manner. In this paper, we propose a framework to model human-human and human-machine interactions based on large-scale real activity data in a bottom-up approach.

3. REPRESENTATION AND INTERPRETATION OF INTERACTION

We introduce four elements of a layered structure model of interaction: primitives, patterns, motifs, and pattern clusters as shown in Figure 2. The lower layer represents a higher abstraction in the figure. They are used for coding and summarizing various kinds of interaction such as dialogs and meetings. Mase et al. [1] proposed primitive, pattern, and complex pattern as the elements. We newly introduce *motif* [5] and *pattern cluster* to handle sequential interaction processes and the meaning of occurring elemental events. In the experiment described later, we will try to summarize medical dialog in terms of pattern clusters and characteristic patterns, which will be defined in the following subsections with the given experimental data.

3.1 Hierarchical Indices

Primitive. A primitive is a basic element of interaction such as "SPEAK" and "GAZE". For example, it represents such an occurring event as "Person A speaks" and "Person A gazes at Person B." A primitive is denoted by Pri in the following discussion.

Pattern. A pattern, a set of primitives that occurs simultaneously, is denoted by Pat.

Motif. A motif, a continuous sequence of multiple patterns that appears more than once, is denoted by Mot in the following discussion: $Mot = (Pat_1, Pat_2, ..., Pat_n)$ shows a motif where patterns appear from Pat_1 to Pat_n in the subscript order. We exclude sequences that appear only once from the final motif because it generates many meaningless motifs.

Pattern Cluster. We define that the patterns contained in a motif are co-occurrence patterns if the motif occurs occasionally. We later introduce a novel pattern clustering algorithm by such a co-occurrence measure given by the motif evaluation. We can describe interactions with fewer labels in terms of pattern cluster.

We omit the start and end times of each element for simplicity of description.

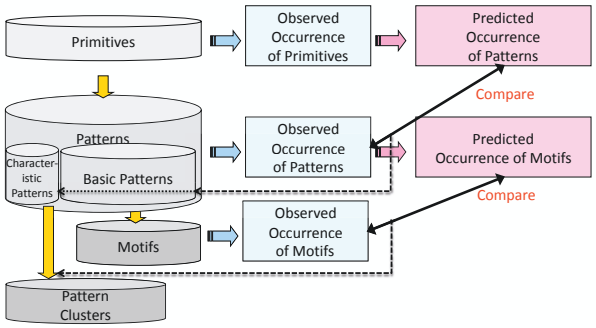

Figure 2: Hierarchical representation of interactions (Observed and estimated occurrences are compared to chose target patterns.)

3.2 Measure of Identification

The possible number of kinds of patterns and motifs defined in the previous section becomes combinatorially very large. To make the resulting corpus useful, we need to reduce the number of usable index patterns and motifs by defining some measures of importance.

There are at least four such measures in text mining: comprehension, eccentricity, identification, and representation. Generally a comprehension measure is rated highly for higher frequency of text, while an eccentricity measure is rated highly for lesser frequency of text. The normalized occurrence measure proposed in [6] is an example of an eccentricity measure. We need to use a measure with a balanced integration of two extreme attributes. One is "tf-idf", for example, an empirically obtained representative measure. The other is Kullback-Leibler Divergence (KLD) [7], which we exploit in this paper. It is based on information theory and can give a balanced identification measure.

We define the importance of pattern and motif by comparing actual event occurrence probability p to its estimation q. KLD is chosen to obtain descriptive indices for summarization of interaction without losing characteristic patterns. Tf-idf has a tendency to lose some characteristic patterns. Actually we exploit the Jensen-Shannon Divergence (JSD)[8], which is obtained from KLD and symmetric, while KLD is asymmetric. The signed-JSD measure $sjsd$ is defined as follows:

$$sjsd(p,q) = \begin{cases} jsd(p,q) & p \geqq q, \\ -1 \times jsd(p,q) & p < q. \end{cases}$$

3.3 Evaluation of Patterns

We define two evaluation measures of patterns: **basic pattern measure** and **characteristic pattern measure**. The basic pattern measure is computed as a signed-JSD

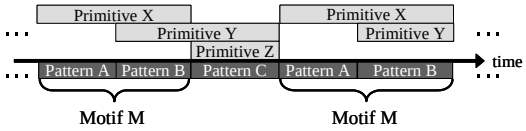

Figure 3: Interval relationships among primitive, pattern, and motif

between the actual occurrence of pattern Pat and the expected occurrence in dataset $D = \{d_1, d_2, \ldots, d_n\}$. The characteristic pattern measure is computed as a signed-JSD between the actual occurrences of pattern Pat in data d and in dataset $D = \{d_1, d_2, ...d_n\}$, where data d are collected from an event sequence, such as a dialog session.

3.3.1 Basic Pattern Measure

First, we define the basic pattern measure $IBP_D(Pat)$ of the pattern Pat in dataset D as the following:

$$IBP_D(Pat) = \sum_{d \in D} \frac{T_d \times sjsd(P_d(Pat), Q_d(Pat))}{\sum_{d \in D} T_d}.$$

where $P_d(Pri)$ is the occurrence of primitive Pri in d and $Q_d(Pat)$ is the expected occurrence of the pattern Pat in data d by the following equation:

$$Q_d(Pat) = \prod_{Pri_n \in Pat} P_d(Pri_n) \prod_{Pri_m \notin Pat} P_d(Pri_m),$$

We can extract the basic patterns by thresholding the basic pattern measure: $IBP_D(Pat) > Th_{IBP}$. The obtained basic patterns are considered suitable to describe overall dataset D.

3.3.2 Characteristic Pattern Measure

Similarly, we define the characteristic pattern measure $ICP_D(Pat)$ over D:

$$ICP_D(Pat) = Max(sjsd(P_d(Pat), P_D(Pat))|d \in D\}).$$

where $P_D(Pat)$ is the average occurrence of the pattern Pat in dataset D:

$$P_D(Pat) = \sum_{d \in D} \frac{T_d \times P_d(Pat)}{\sum_{d \in D} T_d}.$$

We can extract the characteristic patterns by thresholding the characteristic pattern measure: $ICP_D(Pat) > Th_{ICP}$.

3.4 Evaluation of Motif

$IBM_D(Mot)$, the motif extraction measure over dataset D, is given as the following:

$$IBM_D(Mot) = \sum_{d \in D} \frac{T_d \times sjsd(P'_d(Mot), Q'_d(Mot))}{\sum_{d \in D} T_d},$$

where $Q'_d(Mot)$ and $P'_d(Mot)$ are the normalized expected occurrence and the normalized actual occurrence of motif, respectively. We can extract **basic motifs** by thresholding the motif extraction measure.

3.5 Pattern Clustering

We exploit the motif evaluation result to define the distance between the patterns in terms of similarity. We assume that patterns are similar when they co-occur consecutively.

The distance between patterns $Dist(Pat_n, Pat_m)$ can be given by the inverse of the pattern co-occurrence measure derived by the motif measure. We can use any appropriate clustering technique since we obtain a distance measure between patterns. We will use a conventional Ward method to cluster the patterns in our experiment.

4. EXPERIMENTAL RESULTS

4.1 Simulated Interview Dataset

We videotaped ten simulated medical interviews at Nagoya University Hospital. The simulated interviews took place with volunteer simulated patients (SP) for the training of the interview skills of medical students. Videotaping and group review conferencing are standard lessons. Each interview lasted about 10 minutes. We digitized the video and manually annotated the dialog with an annotating tool we developed for this purpose at the 0.1 second precision. 12 kinds of frequently observed primitives were chosen, such as speak, gaze to human, gaze to memo, head nod, rhythm, and touching self for SP. The same 12 kinds and a memo-taking primitive were chosen for the doctor-role medical students. A total of 25 kinds of primitives were chosen at last based on the physician's advise as well as the listing of major nonverbal behaviors in communication psychology research literatures. The chosen primitives sensing can be automated by the recent vision and speech technologies.

4.2 Extracted Patterns and Motifs

The number of observed patterns is 1,569. First, we applied threshold $Th_{IBP} = 0.001$ for the basic pattern measure and extracted the top 18 patterns out of 1,569. The average total time with the 18 patterns covered 45% of the interview times. For example, the top pattern is where one person is speaking while eye contact is being made. Next, we applied another threshold, $TH_{ICP} = 0.003$, to extract the 14 characteristic patterns that cover 7% of the interview time. These thresholds are chosen based on observation of each measure distributions. One example of a characteristic pattern is where a patient is touching his/her body when explaining a symptom. This pattern appeared specifically in a few interviews. Third, we obtained 900 motifs with 18 extracted basic patterns. 13 basic motifs were extracted out of 900 by thresholding with TH_{IBM}.

4.3 Pattern Cluster Extraction

Pattern clusters were extracted based on our proposed method described in Section 3.5. We analyzed the 18 basic patterns and observed three clusters or six sub-clusters (two sub-clusters per cluster) using a Ward method clustering as shown in Figure 4. The clusters were interpreted as follows: (1) memo taking, (2) medical student utterances, and (3) mostly patient utterance related events. The sub-clusters were (1-a) memo taking and joint gaze to memo, (1-b) only memo taking and so on.

4.4 Summary Reading Experiment

We visualized how six pattern clusters and 14 characteristic patterns occur in sequence for each interview. The pattern cluster indices reveal the overall dialog flow, while the characteristic patterns emphasize the specific situation of each interview. We asked a physician to read the ten visualized summaries and to make notes about his reading and

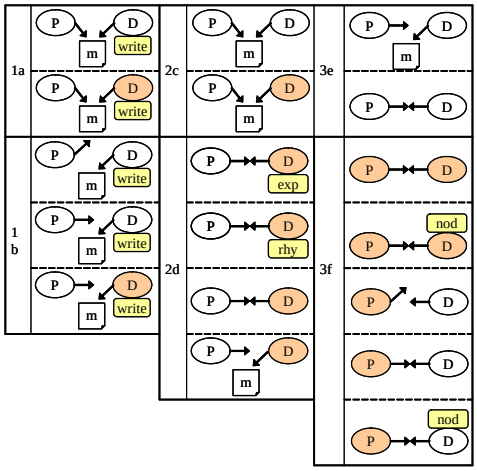

Figure 4: Pattern clusters(P: patient, D: doctor, arrow: eyegaze, hatch/orange: utterance, m: memo, nod: nodding, rhy: rhythmic gesture, exp: iconic explaining, dir: deictic to affected part, write: memo writing)

impressions. We then asked him to review the interview videos to compare the impressions from videos and written notes to give one of four answers: (a) matched features, (b) mis-matched features, (c) unknown from summary, and (d) other.

The following are the comparison results between the summaries and video reviews: (a) matched features: 39.5%, (b) mis-matched features: 26.3%, (c) unknown from summary: 26%, and (d) other: 13.2%. The total number of comparison items is 38. 39.5% matches is a fairly good result. The following was the physician's overall impression: *I expected that the Doc would allow SP to speak sufficiently and bring the interview to the Doc's pace. This was confirmed by the video. The example was very successful.*

5. CONCLUSION

We introduced four layered indices, primitive, pattern, motif, and pattern cluster to describe general interaction between doctors and patients in a bottom-up manner. We proposed a bottom-up method of indice generation for a corpus. It used a data set to include various possible pattern events for consideration and to reduce them to representative, higher level notations. Jensen-Shannon Divergence was exploited to give these measures and to extract important and key patterns to summarize dialogs, for example. We applied the proposed method to medical interview sessions between medical students and simulated patients. The experimental results show a good potential of the approach. This kind analysis will help tutors and students to identify the strength and weakness of each student's interview skill quantitatively. The annotated video will be used more efficiently in the review session to access to the point of interest. We hope, in the future, the physicians can review his/her recorded interviews as they do for medical operations.

There are some remaining issues to solve in the future. First, the size of data set is not large enough so as to claim that the obtained patterns, motifs and clusters are general enough to express the medical interview in general. However, we think the proposed method will lead a fairly good results when we will prepare a large scale data set. Second, there are a few heuristic thresholding parameters in the index selection as criteria to obtain appropriate results. The choice of parameter is a kind of design how we want to describe the dialog in detail. Third, we only exploited motifs for pattern cluster evaluation in this paper. In the future we will extend our method to generate motif segments to improve our summary. That would lead to a higher level description of interview and we may be able to extract any physician's strategy of dialog in medical interview. We also want to extend the description by adding attributes such as the strengths of eye gaze and hand gesture movement.

The proposed method can be applied to general dialog scenes and interactions. We look forward to extend toward wider domains; e.g. multi-party dialog and sports scene. In a sport application, for example, if primitive plays are annotated properly, we can extract major patterns and motifs including point getting events, with which we will be able to obtain abstract documentation of a game with hierarchical indices. Based on the proposed bottom up approach, we have more chances to come across the important moves/plays which have been overlooked by conventional coaching, score-book, and commentary.

Acknowledgment

This research is supported in part by a JSPS Grant-in-Aid (18300048) and by National Institute of Information and Communications Technology (NICT). This research was approved as a clinical test research by the Ethics Review Committee, Nagoya University Medical School (number: 442).

6. REFERENCES

[1] K. Mase et. al. : Ubiquitous Experience Media, *IEEE Multimedia*, Vol. Oct-Dec, pp.20-29 (2006).

[2] C. R. Wren, Y. A. Ivanov, I. Kaur, D. Leigh and J. Westhues: SocialMotion: Measuring the Hidden Social Life of a Building, Location- and Context- Awareness 2007, *LNCS*, Vol. 4718, pp. 85-102 (2007).

[3] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita, Interactive Humanoid Robots for a Science Museum, IEEE Intelligent Systems, Vol. 22, No. 2, pp. 25-32, Mar/Apr, 2007.

[4] K. Otsuka et. al., A Probabilistic Inference of Multiparty-Conversation Structure Based on Markov-Switching Models of Gaze Patterns, Head Directions, and Utterances, ICMI'05, pp.191-198, October, 2005.

[5] J. Lin, E. Keogh, S. Lonardi, and P. Patel: Finding Motifs in Time Series, In the 2nd Workshop on Temporal Data Mining, pp. 23-26 (2002).

[6] T. Morita, Y. Hirano, Y. Sumi, S. Kajita, and K. Mase, A Pattern Mining Method for Interpretation of Interaction, ICMI'05, pp. 267-273, Oct. 2005.

[7] S. Kullback and R.A.Leibler: On Information and Sufficiency, *The Annals of Mathematical Statistics*, Vol. 22, No. 1, pp. 79-86 (1951).

[8] B. Fuglede, and F. Topsoe: Jensen-Shannon Divergence and Hilbert space embedding, *Proceedings of the International Symposium on Information Theory*, pp. 31-36 (2004).