Usage Patterns and Latent Semantic Analyses for Task Goal Inference of Multimodal User Interactions

Pui-Yu Hui, Wai-Kit Lo and Helen M. Meng
Human-Computer Communications Laboratory
The Chinese University of Hong Kong
Shatin, Hong Kong SAR of China
{pyhui, wklo, hmmeng}@se.cuhk.edu.hk

ABSTRACT

This paper describes our work in usage pattern analysis and development of a latent semantic analysis framework for interpreting multimodal user input consisting speech and pen gestures. We have designed and collected a multimodal corpus of navigational inquiries. Each modality carries semantics related to domain-specific task goal. Each inquiry is annotated manually with a task goal based on the semantics. Multimodal input usually has a simpler syntactic structure than unimodal input and the order of semantic constituents is different in multimodal and unimodal inputs. Therefore, we proposed to use semantic analysis to derive the latent semantics from the multimodal inputs using latent semantic modeling (LSM). In order to achieve this, we parse the recognized Chinese spoken input for the spoken locative references (SLR). These SLRs are then aligned with their corresponding pen gesture(s). Then, we characterized the cross-modal integration pattern as 3-tuple multimodal terms with SLR, pen gesture type and their temporal relation. The inquiry-multimodal term matrix is then decomposed using singular value decomposition (SVD) to derive the latent semantics automatically. Task goal inference based on the latent semantics shows that the task goal inference accuracy on a disjoint test set is of 99%.

Author Keywords

Multimodal input, spoken input, pen gesture, task goal inference, latent semantic modeling (LSM), singular value decomposition (SVD).

ACM Classification Keywords

H.5.2 [User Interfaces]: Input devices and strategies, Natural language, User-centered design.

1. INTRODUCTION

This paper describes our initial attempt in developing a semantic analysis framework for multimodal user input with speech and pen gestures. Each modality in the multimodal user input presents a different abstraction of user's informational or communicative goal as one or more input events. Semantic interpretation of multimodal inputs captured with the mobile devices requires syntactic, semantics, temporal and contextual information derived from multiple modalities. Previous work in semantic interpretation of multimodal input include frame-based heuristic integration [1, 2], unification parsing [3, 4], hybrid symbolic-statistical approach [5, 6], weighted finite-state transducers [7], probabilistic graph matching [8, 9] and the salience-driven approach [10]. We leveraged such experiences to devise a computationally efficient approach based on score-based, cross-modal integration that incorporates semantic and temporal information [11]. The approach does not present high demands for training data. The current work extends cross-modal integration with semantic interpretation. More specifically, our aim is to infer the domain-specific task goal(s) of the multimodal input. The task goal is characterized by terms used in the spoken modality, as well as particular term co-occurrence patterns across modalities. Previously, we have applied Belief Networks [12, 13] for task goal inference based on unimodal (speech-only) inputs. However, previous studies [11, 19] that compare the spoken part of multimodal inputs with unimodal (speech-only) inputs shows that the former generally has simpler syntactic structures, more diverse vocabularies and different term ordering. Therefore, we explore the use of latent semantic modeling (LSM) for task goal inference, with the objective of uncovering the associations between (unimodal or multimodal) terms and task goals through a data-derived latent space.

LSM [14] is a data-driven approach that models the underlying semantics of word usages from available corpora. It has been applied unimodally to text or transcribed speech for language modeling [15], document clustering [16], spoken document retrieval [17], document summarization [18], etc. The objective of our current work is to apply LSM in capturing the latent semantics of the multimodal user inputs. In LSM, the associations of between terms and inquiries are represented as a term-

IUI'10, February 7-10, 2010, Hong Kong, China.
Copyright 2010 ACM 978-1-60558-515-4/10/02…$10.00.

inquiry matrix. This can be factorized into a term-semantics and an inquiry-semantics matrix using singular value decomposition (SVD) [14]. These two matrices associate terms and inquiries through an automatically derived space of semantics, instead of directly relating the terms with inquiries.

We represent a multimodal input by means of lexical or multimodal terms. Multimodal terms are decided base on cross-modality integration patterns elicited from the user inputs. We then perform LSM to analyze the content of a multimodal input. Each input is associated with every latent semantic category by a weight. The weights are used for task goal inference. There are a total of nine task goals in our experimental domain.

In the following, we introduce latent semantic analysis, present the collected multimodal corpus and discuss the process of task goal inference and related experimentation.

2. THE LATENT SEMANTIC ANALYSIS FRAMEWORK

We apply latent semantic analysis for task goal inference based on multimodal input. The latent semantic model (LSM) uses Singular Value Decomposition (SVD) to derive a latent semantic space that relates multimodal terms (combined lexical and gestural terms) with the users' inputs. Correlations between cross-modal terms are captured from the training data. During testing, multimodal terms are extracted from the input and the vector is projected into the latent space. Thereafter, the task goal is inferred based on a combination of latent semantics.

2.1 Association Matrices

Associations between terms and inquiries can be summarized in a term–inquiry matrix G. Given M terms (details of the multimodal terms will be presented in section 4.5) and N inquiries, we form an $M \times N$ matrix G. Each column represents an inquiry. The element $g_{\mathrm{m,n}}$, is the weight (i.e. normalized term frequency using TF-IDF) for the term m in the n^{th} inquiry.

$$G = \begin{bmatrix} g_{1,1} & \cdots & g_{1,n} & \cdots & g_{1,N} \\ \vdots & \ddots & \vdots & \cdot^{\cdot^{\cdot}} & \vdots \\ g_{m,1} & \cdots & g_{m,n} & \cdots & g_{m,N} \\ \vdots & \cdot^{\cdot^{\cdot}} & \vdots & \ddots & \vdots \\ g_{M,1} & \cdots & g_{M,n} & \cdots & g_{M,N} \end{bmatrix} \qquad (1)$$

where $g_{m,n} = (1-\varepsilon_m)\dfrac{\kappa_{m,n}}{\lambda_n}$,

$$\varepsilon_m = -\frac{1}{\log N}\sum_{n=1}^{N}\frac{\kappa_{m,n}}{\tau_m}\log\frac{\kappa_{m,n}}{\tau_m},$$

$\kappa_{m,n}$ denotes the number of times the term m occurs in the n^{th} inquiry,

λ_n is the total number of terms in the n^{th} inquiry,

ε_m denotes the normalized entropy of term m in the data set; and

τ_m is the total number of times that term m occurs in the training set.

G may be decomposed into a product of three matrices, with methods such as singular value decomposition (SVD), probabilistic latent semantic analysis (PLSA) [20] and latent Dirichlet allocation (LDA) [21]. We propose to focus on the use of SVD of order R, as shown in Equation 2.

$$G = USV^{T}$$
$$= \begin{bmatrix} u_{1,1} & \cdots & u_{1,R} \\ \vdots & \ddots & \vdots \\ u_{M,1} & \cdots & u_{M,R} \end{bmatrix} \begin{bmatrix} s_{1,1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_{R,R} \end{bmatrix} \begin{bmatrix} v_{1,1} & \cdots & v_{1,R} \\ \vdots & \ddots & \vdots \\ v_{N,1} & \cdots & v_{N,R} \end{bmatrix}^{T} \qquad (2)$$

where U is the left unitary matrix of dimensions M x R,

S is the diagonal matrix of singular values sorted in descending order with dimensions R x R,

V^T is the right unitary matrix of dimensions R x N,

R=min{M, N} is the order of decomposition and

T is the transpose of the matrix.

Each column of U contains the estimated weight of each term m that corresponds to the latent semantic category r while each column of V^T contains the estimated weight of each inquiry n that corresponds to the latent semantic category r. Equation 2 projects the space of terms and inquiries onto a reduced R-dimensional space which is defined by the orthonormal basis given by the column vectors u_m and v_n from matrices U and V respectively. In order to collapse the terms that are "semantically similar", we always choose $R'<R$. The smaller the value R', the more pronounced is the reduction of semantic redundancy in the latent semantic space. Based on the latent semantic space, we may re-construct the space of terms and inquiries, denoted as $\hat{G}$ in Equation 3.

$$G \approx \hat{G} = U\hat{S}V^{T} \qquad (3)$$

where $\hat{S}$ is the reduced diagonal matrix of singular values with optimized value of R'.

We need to find an "optimal" choice of R' that minimizes the distortion between the re-constructed space $\hat{G}$ and the original space G, in the implementation of Equation 3 in the training procedure. We plan to optimize R' through empirical analysis of the latent space.

2.2 Relating Task Goals with Latent Semantics

In the training procedure, we represent the n^{th} inquiry by the column vector g_n. The weights for latent semantic category r can then be obtained by a dot product between g_n and the corresponding column vector of the left unitary matrix U, u_r. Therefore, from the vector g_n, we can obtain a vector of weights w_n for each latent semantic category by Equation 4:

$$w_n = g_n^{T}U \qquad (4)$$

where $w_n = \begin{bmatrix} w_{n,1} & \cdots & w_{n,R'} \end{bmatrix}$, $g_n = \begin{bmatrix} g_{1,n} & \cdots & g_{M,n} \end{bmatrix}^{T}$ and

$w_{n,r}$ is the weight of latent semantic category r for the n^{th} inquiry.

We use A to denote the total number of task goals within the application domain, a_n to denote the task goal of the n^{th} inquiry, and R' to denote the number of dimensions in the latent semantic space. We attempt to compute a projection matrix F that can transform the vector of weights for the latent semantic categories of inquiry n (i.e. w_n) into a vector of weights for the A task goals (see Equation 5).

$$h_n = w_n F \tag{5}$$

where $F = \begin{bmatrix} f_{1,1} & \cdots & f_{1,A} \\ \vdots & \ddots & \vdots \\ f_{R',1} & \cdots & f_{R',A} \end{bmatrix}$,

$f_{r,a}$ is the weight of a latent semantic category r that would correspond to a task goal a and

$h_n = \begin{bmatrix} h_{n,1} & \cdots & h_{n,A} \end{bmatrix}$

where $h_{n,a}$ is the weight of the n^{th} inquiry would correspond to a task goal a.

According to Equation 5, associations between inquiry and latent semantic categories can be summarized in an inquiry-latent semantic categories matrix W (an NxR' matrix) and the associations between inquiry and task goal can be summarized in an inquiry-task goal matrix H (an NxA matrix). Therefore, we can obtain Equation 6 as follows.

$$H = WF \tag{6}$$

where $W = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} w_{1,1} & \cdots & w_{1,R'} \\ \vdots & \ddots & \vdots \\ w_{N,1} & \cdots & w_{N,R'} \end{bmatrix}$ and

$$H = \begin{bmatrix} h_1 \\ \vdots \\ h_N \end{bmatrix} = \begin{bmatrix} h_{1,1} & \cdots & h_{1,A} \\ \vdots & \ddots & \vdots \\ h_{N,1} & \cdots & h_{N,A} \end{bmatrix}.$$

Mathematically, the projection matrix F can be found using Equation 7.

$$F = W^{-1}H' \tag{7}$$

where $H' = \begin{bmatrix} h'_1 \\ \vdots \\ h'_N \end{bmatrix} = \begin{bmatrix} h'_{1,1} & \cdots & h'_{1,A} \\ \vdots & \ddots & \vdots \\ h'_{N,1} & \cdots & h'_{N,A} \end{bmatrix}$,

h'_n is a vector of manually labeled task goal for n^{th} inquiry,

$h'_{n,a}$ is the manually labeled task goal of inquiry n, in which $h'_{n,a}=\{0,1\}$ and $\sum_{a=1}^{a=A} h'_{n,a} = 1$ and

W^{-1} is the pseudo inverse of the matrix W.

Through the projection matrix F and Equation 5, we can obtain the weight of each inquiry that would correspond to each task goal. A task goal a_n^* will be assigned as the automatic derived task goal for inquiry n where $a_n^* = \arg\max_a \{h_{n,a}\}$.

The performance of task goal inference of the training data can then be evaluated by comparing a_n^* to the manually annotated task goal a_n. Moreover, we may examine the structural relations between latent semantic category and task goals in the transformation matrix F.

In the testing procedure, we also represent the n^{th} inquiry by a vector g_n. We obtain the weights for the r latent semantic categories by Equation 4 where the left unitary matrix U is obtained from the training procedure. The vector of weights for each latent semantic category lies in the R'-dimensional space. We transform it to A-dimensional space and automatically derived task goal a_n^* for the n^{th} inquiry using Equation 5. The task goal inference performance can be evaluated by comparing the a_n^* assigned and task goal a_n manually annotated of the n^{th} inquiry.

3. MULTIMODAL CORPUS

We collected a multimodal corpus with speech and pen gestures in the city navigation domain. We invited 23 Mandarin-speaking subjects, each of whom was asked formulate 66 task-oriented multimodal inputs according to a set of instructions. The tasks are designed based on nine task goals including:

- BUS_INFORMATION
- CHOICE_OF_VEHICLE
- MAP_COMMANDS
- OPENING_HOURS
- RAILWAY_INFORMATION
- ROUTE_FINDING
- TIME_CONSTRAINT
- TRANSPORTATION_COSTS and
- TRAVEL_TIME

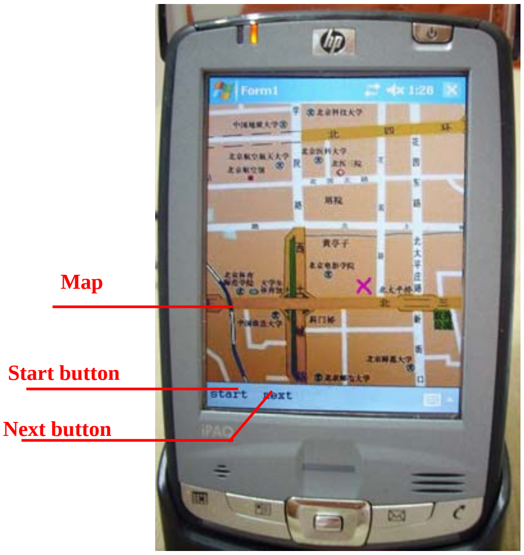


Figure 1. Data collection interface of the Pocket PC, augmented with soft buttons for logging functions (start/stop) and loading the next map.

Each inquiry may involve up to n (n=6 in this work) locations. They may refer to target locations shown on a Pocket PC interface (see Figure 1) by speech and/or pen gestures. Both speech and pen input are recorded directly by the Pocket PC. Captured multimodal inputs involve disfluencies in the speech modality (e.g. filled pauses and repairs), spurious pen gestures (e.g. repetitive pen gesture inputs) and recognition errors in both modalities. These imperfections have adverse effects on cross-modality semantic integration. Table 1 shows an example task and a multimodal input obtained during data collection. We collected a total of 1518 user inquiries, which include 1442 multimodal and 76 speech-only (unimodal) inputs. All speech and pen data are *manually* transcribed. Furthermore, subjects were presented with their own multimodal data and asked to indicate (based on their original intentions) the correspondences between the spoken locative references (e.g. *here*, *the nearest station*, etc.) and pen gestures. These cross-modality correspondences are referenced as we annotate the cross-modal integration patterns. We divided the 1,442 multimodal inputs randomly into disjoint training and test sets in a 7:3 ratio. Hence we have 999 inputs as the training set and 443 inputs as the test set. As a first step, we apply the task goal inference framework to the *manual* transcriptions of speech (which is equivalent to having perfect speech recognition) and will defer handling speech recognition errors to a later step.

Information category: TRAVEL_TIME
Task: 告知系統你所在的位置，查詢從那裡到另外四所大學需要多長時間。 “*Specify your current location. Find the time it takes to travel to four universities of your choice.*”
Multimodal input: ***S***: 我在 北郵 從 這裡 出發到 這四個大學 要多久？ ***P***: • (*a point*) •••• (*four points*) “*I’m at BUPT. From here, I want to visit these four universities. How long will it take?*”

Table 1. Example of a multimodal input with speech (*S*) and pen gestures (*P*). Translations are provided in italics and corresponding SLRs are underlined.

4. CHARACTERIZING AND ANNOTATING CROSS-MODAL INTEGRATION PATTERNS

A navigational inquiry in the multimodal corpus may include one or more SLRs and/or pen gestures that indicate specific locations on the map. As will be explained later, there is no straightforward correspondence between SLRs and pen gestures. Therefore, we need to understand the characteristics of each modality and their temporal relationships, in order to appropriately obtain cross-modal integration patterns.

4.1 Spoken Locative References

An SLR can be a direct (full name, abbreviated name or a contextual phrase such as “*my current location*”) or an indirect one [11]. It may also be a singular, aggregated, plural reference or unspecified on number:

- A **singular reference** can be a direct reference with a full name or an abbreviated name. It may also be a singular indirect reference (e.g. 這個公園 “*this park*”), which may optionally include information about the location type (i.e. a park in the given example).
- An **aggregated reference** is an indirect reference with a specific numeric value (which is greater than 1) and an optional location type feature (e.g. 這四個地方 “*these four locations*”).
- A **plural reference** is an indirect reference with the numeric feature set to plural (i.e. NUM=*plural*), as well as an optional location type feature (e.g. 這些大學 “*these university*”).
- An **unspecified reference** is an indirect reference with unspecified numeric and location type features (e.g. 這裡 “*here*”).

Analysis of the training set shows that it contains 2442 SLRs among the 999 multimodal inputs. Their distribution is shown in Figure 2.

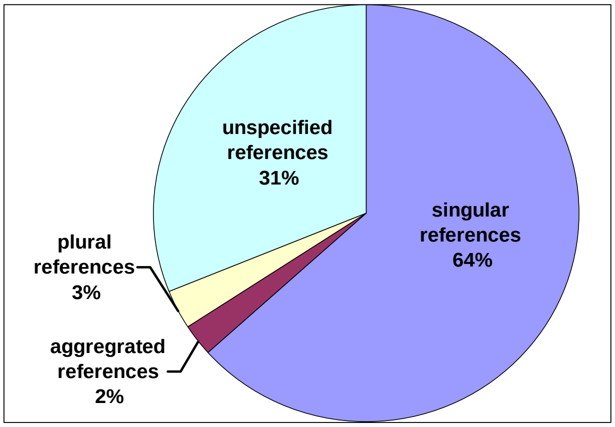


Figure 2. Distribution of the types of SLRs in the training set.

Spoken Terms Regularization

Analysis of the spoken inputs also shows that there are many synonymous terms and aliases. For example, the word “route” in Chinese consists of two characters (i.e. 路線), which may also be reversed (as 線路) and the meaning of the word remains the same. Similarly, SLRs may have synonymous terms. For example, the full name 北京郵電大學 (i.e. *Beijing University of Post and Telecommunications*) may be abbreviated as 北郵 (i.e. *BUPT*). There is also a variety of verbalization to express the contextual phrase of “current location”, including: 目前的所在地, 當前的位置, 所在的地方, 所在地, etc. Other contextual phrases may differ by a “measure word” which is characteristic of Chinese, e.g. 這間大學 and 這所大學 both mean “*this university*”. In order to simplify processing, synonymous terms and aliases are collapsed into a single category. In other words, we have created a category for each group of semantically equivalent terms. It is conceivable that this categorization may be implemented through the use of SVD if sufficient data is available. Since we only have limited training data for the time being, we choose to design regularization rules (56 rules in all) for categorization.[1] As such, we have reduced

[1] This step forms equivalence classes that group together terms with the same meaning. We expect that this step should help task goal inference because it reduces term diversity given the limited amount of training data. We plan to perform an analogous experiment without term regularization for comparison.

the number of lexical terms[2] significantly. Since we also have pen gestures with their corresponding SLRs, we are able to form "multimodal terms". Each is a 3-tuple consisting of an SLR, the corresponding pen gesture and their temporal relationship. We will elaborate on this later.

The statistics of the lexical and multimodal terms in the training set are shown in Table 2. After regularization, the number of multimodal terms can be reduced by around 66%. The number of (SLR and pen) multimodal terms is fewer than expected. There are 22 multimodal terms that contain only an SLR with no pen gesture. This is because of an anaphoric reference (which can be resolved with contextual information). There are also 6 multimodal terms that contain pen gestures only and no SLR, due the use of ellipsis.

	Before term regularization	After term regularization
# of Multimodal terms (SLR and pen) (SLR only) (Pen only)	456 407 43 6	261 233 22 6
# of Lexical terms	260	216
Total number of terms	716	477

Table 2. Statistics of the lexical and multimodal terms (count by type).

4.2 Pen Gestures

A pen gesture can be recognized as a point, circle or stroke. Using these three types of pen gestures, subjects can indicate different semantics such as a single location, an area with multiple locations or a route (see Table 3). There can be up to *n* (=6) locations in an inquiry and the mapping between SLRs and pen gestures may be one-to-many or many-to-one. Analysis of the training set shows that it contains 2480 pen gestures. 95% of the multimodal inputs contain a single pen gesture, i.e. POINT, CIRCLE or STROKE. The remaining multimodal inputs (i.e. 5%) contain multiple pen gestures, to which we refer as MULTI-POINT, MULTI-CIRCLE and MULTI-STROKE. Table 3 shows examples of pen gestures and their semantics.

4.3 Correspondence between Spoken Locative References and Pen Gestures

We derive the correspondence between SLRs and pen gestures based on temporal ordering and semantic compatibility. Analysis of the training data shows that in a multimodal input, SLR and pen gesture that (jointly) refer to the same intended location may not always overlap in time. Hence our approach only enforces the temporal ordering of SLRs and pen gestures when deriving their associations.

Additionally, the association between SLRs and pen gestures also enforce semantic compatibility. Our approach checks the numeric value (NUM) of an SLR and ensures that it is associated with a compatible number of pen gestures. For the case of one-to-many mapping between the SLR and its associated pen gestures, the pen gestures are considered together as a group (i.e. MULTI-POINT, MULTI-CIRCLE or MULTI-STROKE) in cross-modality integration. The reverse is also true when mapping a pen gesture to multiple SLRs. Furthermore, our approach also checks for agreement in the location type features (LOC_TYPE) in the cross-modality association.

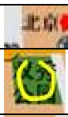
Semantics	Gesture type	Illustration(s)
A single location	POINT	
	CIRCLE	
	STROKE	
An area / multiple locations	CIRCLE (a big circle)	
Multiple locations	MULTI-POINT (four points correspond to one SLR)	
	MULTI-CIRCLE (four circles correspond to one SLR)	
	MULTI-STROKE (three strokes correspond to one SLR)	
A route	STROKE (a stroke indicates the start and end points)	
	MULTI-STROKE (a long stroke with one or more turning points indicate a route)	

Table 3. Examples of different pen gesture types and their semantics.

4.4 Temporal Relationships

Temporal integration patterns [19] between corresponding SLRs and pen gestures, as observed in our training set, include two main types: simultaneous (SIM) and sequential patterns (SEQ). Simultaneous SLRs and pen gestures have temporal overlap. Sequential associations do not. A 3-tuple that consists of corresponding SLR(s) and pen gesture(s), together with their temporal relationship, i.e., <SLR | pen_gesture_type | temporal_relationship> is referred as a multimodal term. Among the 2261 multimodal terms found in the training set, 74% are simultaneous and 26% are sequential. For example, consider the multimodal expression:

[2] A lexical term refers to a tokenized Chinese word from the speech modality but which is not an SLR. Examples include: 開放時間 "*opening hours*", 路線 "*route*", 從 "*from*", etc.

從 我所在的地方 到 這裡 可以怎麼走？
● ● ● ● ●

“*How can I go from my current location to here?*”
Its multimodal terms include <我所在的地方|POINT |SIM> and <這裡|MULTI-POINT|SIM>.

Speech (as parsed SLR)	**Pen (as transcribed gesture)**	**Temporal Relationship (SIM / SEQ)**	**Count**
Singular (1550/2480, 62.5%)	Single (1417/1550, 91.4%)	SIM (1024/1417, 72.3%)	1024
		SEQ (393/1417, 27.7%)	393
	Multiple (0/1550, 0%)	SIM	0
		SEQ	0
	∅ (133/1550, 8.6%)	∅	133
Aggregated (56/2480, 2.3%)	Single (9/56, 16%)	SIM (7/9, 77.8%)	7
		SEQ (2/9, 22.2%)	2
	Multiple (44/56, 78.6%)	SIM (25/44, 56.8%)	25
		SEQ (19/44, 43.2%)	19
	∅ (3/56, 5.4%)	∅	3
Plural (75/2480, 3%)	Single (21/75, 28%)	SIM (12/21, 57.1%)	12
		SEQ (9/21, 42.9%)	9
	Multiple (54/75, 72%)	SIM (35/54, 64.8%)	35
		SEQ (19/54, 35.2%)	19
	∅ (0/75, 0%)	∅	0
Unspecified (761/2480, 30.7%)	Single (715/761, 94%)	SIM (569/715, 79.6%)	569
		SEQ (146/715, 20.4%)	146
	Multiple (1/761, 0.1%)	SIM (1/1, 100%)	1
		SEQ (0/1, 0%)	0
	∅ (45/761, 5.9%)	∅	45
∅ (38/2480, 1.5%)	Single (34/38, 89.5%)	∅	34
	Multiple (4/38, 10.5%)	∅	4

Table 4. Statistics of cross-modal integration patterns in the training set. There are altogether 2480 multimodal terms (count by token) in total. Among them, 2261 contain both SLR and pen gesture, 181 contain only SLRs and 38 of them contain only pen gestures.

4.5 Cross-Modal Integration Patterns

Recall that SLRs may be singular, aggregated, plural and unspecified references. Recall also that an SLR may correspond to one or more pen gestures. We analyze the statistics in the training set as shown in Table 4. From the statistics, we observe that users predominantly prefer to use a single reference in the SLR (62.5%). Furthermore, a single SLR generally corresponds to a single pen gesture, as none were found mapping to multiple pen gestures. As regards aggregated references (e.g., <這四個大學> or <*these four universities*>), 79% were found to correspond with multiple pen gestures to indicate multiple locations. The other 16% are used with a circle (i.e. a single pen gesture) that encompasses multiple locations. An example is the multimodal term <這四個大學|CIRCLE|SIM> or <*these four universities |CIRCLE|SIM*>. For plural references, 72% are used with multiple pen gestures to indicate multiple locations. The remaining 28% are used with a single pen gesture, with the majority (19/21) being circles and the remaining two are points. SLRs with an unspecified numeric features should correspond to both single and multiple pen gestures. Within the training set, however, an unspecified reference predominantly (94%) occurs in association with a single pen gesture.

User input with deictic and anaphoric references (the second “*here*” is an anaphora to the first “*here*”): *S*: 我 在 這裡 從 這裡 到 這裡 要 多久 *P*: • ○ “*I’m now here. How much time will it take to go from here to here?*”
Annotated user input with multimodal terms: 我 在 <這裡|POINT|SIM> 從 <這裡|∅|∅> 到 <這裡|CIRCLE|SEQ> 要 多久 “*I’m now at <here|POINT|SIM>. How much time will it take from <here|∅|∅> to <here|CIRCLE|SEQ>?*”
User input with elliptic locative references (the SLR is omitted in speech): *S*: 開放時間 “*Opening hours?*” *P*: •••
Annotated user input with a multimodal term: <∅|MULTI-POINT|∅> 開放時間 “*<∅|MULTI-POINT|∅> Opening hours?*”

Table 5. Examples on 3-tuple multimodal term annotation with speech (*S*) and pen gesture (*P*). Translations are italicized and quoted.

The above refers to SLRs that are deictic or anaphoric expressions. Deictic expressions need to be interpreted jointly with the associated pen gestures. Anaphoric references are interpreted based on contextual information and do not correspond to any pen gestures. The first row in Table 5 presents examples of these two types of expressions. Additionally, there are also elliptic expressions, where the SLR is completely omitted but the pen gesture is present. For such cases, the cross-modal temporal relationship is irrelevant (and indicated by “∅”). Table 5 shows some examples.

The number of multimodal terms is much fewer than the exhaustive combinations between SLRs and pen gestures. Some of the terms are not available in the corpus, while others may be implausible combinations, such as:

A singular reference with multiple pen gestures (e.g. <這個大學|MULTI-POINT|SIM> "*<this university |MULTI-POINT|SIM>*") – a singular SLR refers to one location and corresponds to one pen gesture. Multiple pen gestures should correspond to an aggregated or plural reference. Therefore, this combination involves incompatibility in the numeric feature.

An aggregated reference with a single point or a single stroke (e.g. <這三個地方|POINT|SIM> "*<these three places |POINT|SIM>*") – an aggregated SLR refers to multiple locations and should correspond to multiple pen gestures or a circle. Again, this combination involves incompatibility in the numeric feature.

An unspecified reference with multiple circles or strokes (e.g. <這裡|MULTI-STROKE|SIM> "*<here|MULTI-STROKE |SIM>*") – empirically, we have found that about 94% of the unspecified references are used to indicate a single location (as shown in Table 4). A possible reason may be that unspecified SLRs have short durations, during which the subjects may find it difficult to gesture multiple circles or strokes simultaneously.

5. TASK GOAL INFERENCE

In the previous section, we examined the associations between SLRs and pen gestures, leading to the definition of a multimodal term that captures cross-modal integration patterns and their temporal relationships. In this section, we present a framework for inferring the task goal based on an input inquiry.

As a reference baseline, we apply the vector-space model [22] for task goal inference. For each task goal a, we consider all of its training expressions and their multimodal terms. We create a vector j_a of weights, using the normalized term frequency TF-IDF of the multimodal terms. For an input multimodal expression, we create a vector g_n, similar to the column vector of G in Equation 1. The similarity between an inquiry g_n and task goal vector j_a is calculated as the inner product of the two vectors. Long inquiries contain more terms. Since the dot product favors long inquiries by generating higher similarity scores, we apply cosine normalization (see Equation 8) to reduce the adverse effect of term repetition.

$$Similarity_{\text{cosine}}(j_a, g_n) = \frac{j_a \cdot g_n}{\|j_a\|\|g_n\|} \qquad (8)$$

where j_a is the weight for all terms in the a^{th} task goal and g_n is the weight for all terms in the n^{th} inquiry.

The input expression is assigned to the task goal a_n* which has the maximum similarity score, as shown in Equation 9.

$$a_n* = \arg\max_a \left\{ Similarity_{\text{cosine}}(j_a, g_n) \right\} \qquad (9)$$

Experiments show that vector-space model can correctly infer task goals for 85% and 90% of the inquiries in training and test sets respectively.

Recall that the proposed approach using LSM involves setting up a term-inquiry matrix G. We include both lexical (unimodal, speech only) terms and multimodal terms with speech and pen gestures. There are a total of 216 unimodal terms and 261 multimodal terms in our training corpus. Hence the non-negative matrix G (in Equation 1) is of dimensions 477x999. As described in Section 2, we apply SVD to G and factorize it into U, S and V.

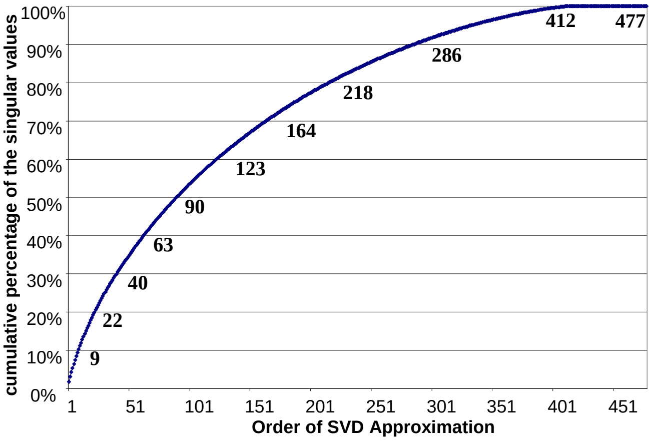


Figure 3. A plot of the cumulative percentage of the singular values against the order of SVD approximation.

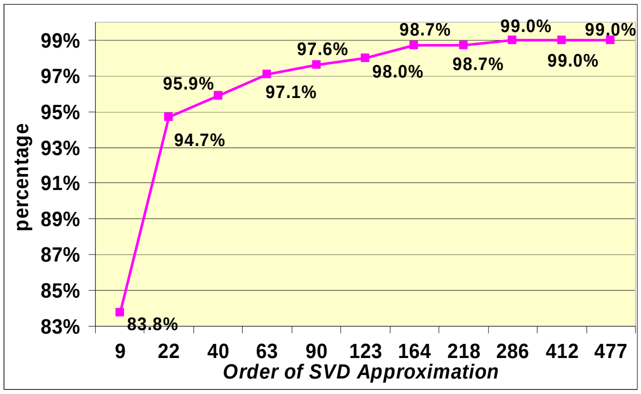


Figure 4. A plot of task goal inference accuracy of multimodal inputs in training set against the order of SVD approximation.

5.1 Optimization of *R'*

Recall that the total number of lexical and multimodal terms sum to R=477. We may consider that the original semantic space to be determined by these terms and attempt to determine the optimal number of dimensions for the latent space. We may choose the order of SVD approximation (R') with reference to the percentage of the cumulative sum of retained singular values over the maximum at R'=477. We plot the percentage of the cumulative sum of preserved singular values over the total sum of all singular values (i.e. at R'=477). In Figure 3, we show the R' values corresponding to the cumulative sum of singular values, at multiples of 10%.

We also perform task goal inference on the multimodal inputs in the training set at the different values of R' (see Figure 4). The performance of task goal inference increases with R'. The rate of increase slowed down as R' becomes higher, reaching saturation approximately at *R'*=286 with a performance of task goal inference at 99% correct. The choice of *R'*=286 as the dimensionality of the latent space implies a reduction of 40% with respect to the original space.

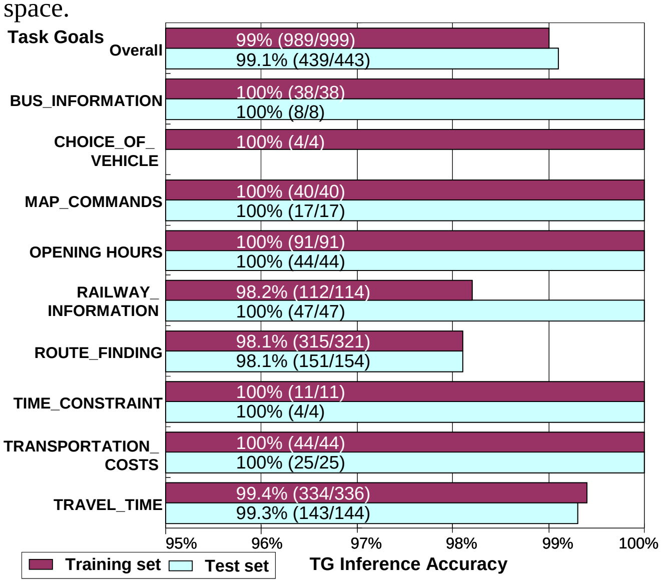


Figure 5. Performance of task goal inference for each of the nine task goals in the application domain. Results are based on the latent space with 286 dimensions.

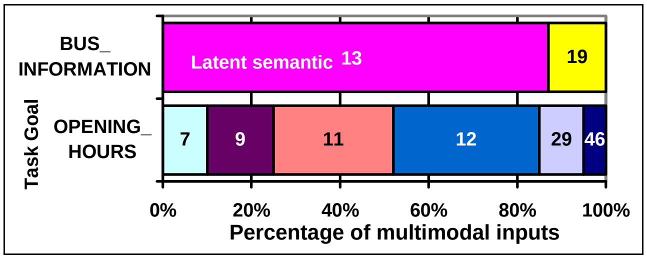


Figure 6. Percentage of multimodal inputs that belong to different latent semantic categories, within two task goals BUS_INFORMATION and OPENING HOURS. The numbers inside the bars are the labels (indexed by *r*) of the latent semantic categories.

5.2 Performance Evaluation

Overall performance in task goal inference for the training and test sets are 99%[3] and 99.1%[4] respectively. Detailed analyses of the results are shown in Figure 5. The test set lacks inqueries that fall under the task goal of CHOICE_OF_VEHICLE (i.e., asking the user what type of vehicle he/she wishes to take). Performance of task goal inference remains high for all the other task goals (at 98% or above).

5.3 Analysis of the Latent Semantic Space

5.3.1 Sub-categorization of task goals

Analysis of the latent semantic space shows that it has sub-divided some of the task goals into logical sub-types. For example, the task goal BUS_INFORMATION contains two latent semantic categories (see Figure 6):

- The latent semantic category (*r*=13) refers to BUS_INFORMATION along a street; e.g. 經過 <這條大街|STROKE|SEQ> 的 所有 公交 線路 是 哪些 "*what are the bus routes that pass through <this street|STROKE|SEQ> ?*"
- The category (*r*=19) refers to BUS_INFORMATION within an area; e.g. 告訴我 所有 在 <這個範圍|CIRCLE|SIM> 行走 的 公交路線 "*please tell me all the bus routes in <this area|CIRCLE|SIM>.*"

Another example is the task goal OPENING_HOURS, which contains six latent semantic categories:

- The latent semantic category (*r*=11) refers to OPENING_HOURS of one location; e.g. 我想知道 <這個公園|POINT|SIM> 的 開放時間 "*I would like to know the opening hours of <this park|POINT|SIM>.*"
- The category (*r*=46) refers to OPENING_HOURS of multiple locations using ellipsis; e.g. <∅|POINT|∅> 開放時間 "<∅|POINT|∅> opening hours."
- The categories (*r*=7 and 29) refer to OPENING_HOURS of multiple locations using multiple singular SLRs; e.g. 我想知道 <這個市場|POINT|SIM> <這個廣場|POINT|SIM> <這個購物中心|POINT|SIM> 的 開放時間 "*I would like to know the opening hours of <this plaza|POINT|SIM>, <this plaza|POINT|SIM> and <this shopping center|POINT|SIM>.*"
- The category (*r*=9) refers to OPENING_HOURS of multiple locations using one aggregated SLR; e.g. 勞駕你告訴我 <這三個地方|MULTI-POINT|SEQ> 的 營業時間 "*Please tell me the opening hours of <these three locations|MULTI-POINT|SEQ>.*"
- The category (*r*=12) refers to OPENING_HOURS of multiple locations using one plural SLR; e.g. 我想知道 <這幾個地方|MULTI-POINT|SEQ> 的 營運時間 "*I would like to know the opening hours of <these locations| MULTI-POINT|SEQ>.*"

We observe that latent semantic modeling has produced subcategories of specific task goals based on the ways in which users compose their inquiries. This is potentially advantageous because finer semantics categorization can enhance understanding and will facilitate automatic generation of system responses.

5.3.2 Capturing key terms for task goals

We examine the term weights in the latent semantic space to identify key terms that are indicative of each task goal. Illustrative examples include:

- For the task goal MAP_COMMAND, key terms with the highest LSM weights are 放大 (i.e. "*zoom in*"), 縮小 (i.e. "*zoom out*"), 拉遠 (i.e. "*zoom out*"), as well as

[3] Improvements in task goal inference accuracies brought about by LSM is statistically significant from 85% to 99% (*α=0.01*, one-tailed *z*-test).

[4] Improvements in task goal inference accuracies brought about by LSM is statistically significant from 90% to 99.1% (*α=0.01*, one-tailed *z*-test).

related standalone pen gestures expressed as the multimodal terms <∅|POINT|∅> and <∅|CIRCLE|∅>.

- For the task goal ROUTE_FINDING, key terms with the highest LSM weights are 到 (i.e. "*to*"), 從 (i.e. "*from*"), 怎樣走 "*how to get to*", 最快 "*the fastest*", 依次 "*in sequence*", as well as the multimodal terms <這裡|POINT|SEQ> (i.e. <*here*|POINT|SEQ>) and <這個大學|POINT|SIM> (i.e. <*this university*|POINT|SIM>).

5.3.3 Generalizing across related multimodal terms

Upon further examination of the LSM weights, we observe their ability to generalize across related multimodal terms, even if the correlations are not directly found in the training data. To describe the underlying mechanism – the LSM framework draws upon the co-occurrences between terms A and B, as well as the co-occurrences between B and C, in order to obtain the correlation between terms A and C.

As an illustration, we can refer to two multimodal inputs by which the user wishes to zoom in on a map:

- 放大 CIRCLE (i.e. the verb phrase "*zoom in*" followed by a circle), corresponding respectively to the lexical and multimodal terms 放大 and <∅|CIRCLE|∅>"
- 放大 POINT (i.e. the verb phrase "*zoom in*" followed by a point), corresponding respectively to the lexical and multimodal terms 放大 and <∅|POINT|∅>"

The column vectors of these two input expressions, as extracted from the original term-inquiry matrix, are shown in Table 7. We compare these vectors with their counterparts in the reconstructed term-inquiry matrix $\hat{G}$ (with *R'*=286), as shown in Table 8. We observe that the reconstructed column vector of the multimodal input "放大 CIRCLE" in Table 8 carry additional weighting (≥0.06) for several additional multimodal terms, namely:

- <這個地方|CIRCLE|SIM>
- <這個範圍|CIRCLE|SEQ>
- <這個範圍|CIRCLE|SIM> and
- <這幅圖|POINT|SIM >

These additional multimodal terms with non-zero weights (see Table 8) did not appear in the original user inputs (see Table 7). But these terms are commly used to convey the task goal MAP_COMMAND, according to the training data (13 out of 40 multimodal inputs). LSM captures the new correlations among <∅|CIRCLE|∅>, 放大 "*zoom in*", <這個地方|CIRCLE|SIM> "*this location*", <這個範圍|CIRCLE|SEQ> "*this area*", <這個範圍|CIRCLE|SIM> "*this area*" and <這幅圖|POINT|SIM > "*this map*" and put them into correlated latent semantics. The weights in Table 8 reflect that the circling action can be used to indicate a single location (這個地方) or an area (這個範圍).

Similarly, we also observe that the feature vector of the multimodal input "放大 POINT in Table 8 introduces additional multimodal terms with non-zero weights (e.g. ≥0.05) for several additional multimodal terms:

- <這個地方|CIRCLE|SIM>
- <這個地方|POINT|SIM> and
- <這幅圖|POINT|SIM >

	放大 "*zoom in*" <∅|*CIRCLE*|∅>	放大 "*zoom in*" <∅|*POINT*|∅>
<∅|CIRCLE|∅>	0.44	0
<∅|POINT|∅>	0	0.34
放大 "*zoom in*"	0.37	0.37
<這個地方|CIRCLE|SEQ> "<*this location*|*CIRCLE*|*SEQ*>"	0	0
<這個地方|CIRCLE|SIM> "<*this location*|*CIRCLE*|*SIM*>"	0	0
<這個地方|POINT|SEQ> "<*this location*|*POINT*|*SEQ*>"	0	0
<這個地方|POINT|SIM> "<*this location*|*POINT*|*SIM*>"	0	0
<這個範圍|CIRCLE|SEQ> "<*this area*|*CIRCLE*|*SEQ*>"	0	0
<這個範圍|CIRCLE|SIM> "<*this area*|*CIRCLE*|*SIM*>"	0	0
<這個範圍|POINT|SIM> "<*this area*|*POINT*|*SIM*>"	0	0
<這個範圍|STROKE|SEQ> "<*this area*|*STROKE*|*SEQ*>"	0	0
<這幅圖|POINT|SIM > "<*this map*|*POINT*|*SIM*>"	0	0

Table 7. An excerpt of the term-inquiry matrix *G* corresponding to two multimodal inputs. The weights (shown up to 2 decimal places) are obtained using Equation 1. Translations are in quotes and italics.

	放大 "*zoom in*" <∅|*CIRCLE*|∅>	放大 "*zoom in*" <∅|*POINT*|∅>
<∅|CIRCLE|∅>	0.18	**0.11**
<∅|*POINT*|∅>	0.06	0.28
放大 "*zoom in*"	0.51	0.44
<這個地方|CIRCLE|SEQ> "<*this location*|*CIRCLE*|*SEQ*>"	0.00	0.00
<這個地方|CIRCLE|SIM> "<*this location*|*CIRCLE*|*SIM*>"	**0.07**	**0.05**
<這個地方|POINT|SEQ> "<*this location*|*POINT*|*SEQ*>"	0.00	0.00
<這個地方|POINT|SIM> "<*this location*|*POINT*|*SIM*>"	0.03	**0.05**
<這個範圍|CIRCLE|SEQ> "<*this area*|*CIRCLE*|*SEQ*>"	**0.07**	0.04
<這個範圍|CIRCLE|SIM> "<*this area*|*CIRCLE*|*SIM*>"	**0.07**	0.04
<這個範圍|POINT|SIM> "<*this area*|*POINT*|*SIM*>"	0.00	0.00
<這個範圍|STROKE|SEQ> "<*this area*|*STROKE*|*SEQ*>"	0.00	0.00
<這幅圖|POINT|SIM > "<*this map*|*POINT*|*SIM*>"	**0.06**	**0.06**

Table 8. An excerpt of the reconstructed term-inquiry matrix $\hat{G}$ corresponding to two multimodal inputs as in Table 7. The estimated weights (shown up to 2 decimal places) of $\hat{G}$ are obtained using Equation 3 with *R'*=286. Translations are in quotes and italics

These additional multimodal terms with non-zero weights (see Table 8) did not appear in the original user inputs (see Table 7). But these terms are commonly used to convey the task goal MAP_COMMAND (11 out of 40 multimodal inputs). LSM captures the new correlations among <∅|POINT|∅>, 放大 *zoom in*, <這個地方|CIRCLE|SIM> "*this location*", <這個地方|POINT|SIM> "*this location*" and <這幅圖|POINT|SIM > "*this map*" and put them into correlated latent semantics.

The weights in Table 8 reflect that the pointing action can be used to indicate a single location (這個地方).

6. CONCLUSIONS

This paper describes our work in the usage pattern and latent semantic analyses of multimodal user inputs with speech and pen gestures. Our investigation is based on a multimodal corpus that we have designed and collected, which consists of over a thousand navigational inquiries. The inquiries cover nine task goals. The task goal of each multimodal input is hand-labeled as a gold standard. We begin with an analysis of the usage patterns and designed the format of a multimodal term to be a 3-tuple, consisting of a spoken locative reference, pen gesture(s) and their temporal relationship). Such multimodal terms can represent the cross-modality integration patterns adopted by the user. Then, we apply latent semantic analysis for task goal inference. Characteristic cross-modal integration patterns are derived from the training set to form multimodal terms. We also derive lexical terms from the speech portion of the multimodal expression. We use a non-negative term-inquiry matrix to capture the associations between terms (lexical and multimodal) and inquiries. Decomposition of the term-inquiry matrix using singular value decomposition captures the associations between terms and inquiries through a latent semantic space. We project the latent semantic space into the space of task goals through a matrix derived from training data. An input multimodal inquiry can be projected into the latent semantic space and then into the task goal space. This gives a vector with which we can use the highest weighting element to select the inferred task goal. We experimented with this approach based on the multimodal corpus. Analysis shows structural relations between latent semantic categories for certain task goals. Furthermore, the weights of the lexical and multimodal terms in the latent semantic space an also help us identify key terms for specific task goals. The latent semantic approach achieves around 99% accuracy in task goal inference, for both the training and test sets. This is significantly higher that the reference baseline obtained with a vector-space model, which achieves 85% and 90% for the training and test sets respectively.

ACKNOWLEDGMENTS

This work is partially supported by a grant from the HKSAR Government Research Grants Council (Project Number 415609). This work is affiliated with the CUHK MoE-Microsoft Key Laboratory of Human-centric Computing and Interface Technologies.

REFERENCES

1. Nigay, L. and J. Coutaz, "A Generic Platform for Addressing the Multimodal Challenge," in the *Proc. of CHI*, 1995.
2. Wang, S. "A Multimodal Galaxy-based Geographic System," S.M. Thesis, MIT, 2003.
3. Johnston, M. et al., "Unification-based Multimodal Integration," in the *Proc. of COLING-ACL*, 1997.
4. Johnston, M., "Unification-based Multimodal Parsing," in the *Proc. of COLING-ACL*, 1998.
5. Wu, L. et al., "Multimodal Integration – A Statistical View," *IEEE Transactions on Multimedia*, 1(4), pp.334-341, 1999.
6. Wahlster, W. et al., SmartKom (www.smartkom.org)
7. Johnston, M. & S. Bangalore, "Finite-state Multimodal Parsing and Understanding," in the *Proc. of COLING*, 2000.
8. Chai, J. et. al., "A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces," in *the Proc. of IUI*, 2004.
9. Chai, J. et. al., "Optimization in Multimodal Interpretation," in *the Proc. of ACL*, 2004.
10. Qu, S. and J. Chai, "Salience Modeling based on Non-verbal Modalities for Spoken Language Understanding," in *the Proc. of ICMI*, 2006.
11. Hui, P. Y. and H. Meng, "Cross-Modality Semantic Integration with Hypothesis Rescoring for Robust Interpretation of Multimodal User Interactions," *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 17, No. 3, 2009.
12. Meng, H., et al., "To Believe is to Understand," in the *Proc. of the Eurospeech*, 1999.
13. Chan, S. F. and H. Meng, "Interdependencies among Dialog Acts, Task Goals and Discourse Inheritance in Mixed-Initiative Dialog," in the *Proc. of the HLT*, 2002.
14. Bellegarda, J. R., "Latent Semantic Mapping: Principles and Applications," *Synthesis Lectures on Speech and Audio Processing*, Vol. 3, No. 1, 2007.
15. Naptali, W., et al., "Word Co-occurrence Matrix and Context Dependent Class in LSA based Language Model for Speech Recognition," *International Journal of Computers*, Issue 1, Volume 3, 2009.
16. Song, W. and S. C. Park, "A Novel Document Clustering Model Based on Latent Semantic Analysis," in the *Proc. of the ICSKG*, 2007.
17. Chen, B., "Word Topic Models for Spoken Document Retrieval and Transcription," *ACM Trans. on Asian Language Information Processing*, Vol. 18, No. 1, 2009.
18. Lee, J. H., et al., "Automatic Generic Document Summarization based on Non-negative Matrix Factorization," *International Journal on Information Processing and Management*, Vol. 45, Issue 1, 2009.
19. Oviatt, S., et al., "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction," in the *Proc. of the CHI*, 1997.
20. Hofmann, T., "Probabilistic Latent Semantic Analysis," in the *Proc. of UAI*, 1999.
21. Blei, D. M., et. al., "Latent Dirichlet allocation," *Journal of Machine Learning Research 3*, 2003.
22. Salton, G. and M. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hall, New York, New Jersey, USA, 1983.