New Perspectives on Spoken Language Understanding: Does Machine Need to Fully Understand Speech?

Tatsuya Kawahara

Kyoto University, ACCMS
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp

***Abstract*—Spoken Language Understanding (SLU) has been traditionally formulated to extract meanings or concepts of user utterances in the context of human-machine dialogue. With the broadened coverage of spoken language processing, the tasks and methodologies of SLU have been changed accordingly. The back-end of spoken dialogue systems now consist of not only relational databases (RDB) but also general documents, incorporating information retrieval (IR) and question-answering (QA) techniques. This paradigm shift and the author's approaches are reviewed. SLU is also being designed to cover human-human dialogues and multi-party conversations. Major approaches to "understand" human-human speech communication and a new approach based on the lister's reactions are reviewed. As a whole, these trends are apparently not oriented for full understanding of spoken language, but for robust extraction of clue information.**

I. Introduction

Automatic speech understanding or spoken language understanding (SLU) has been generally regarded as a next step of automatic speech recognition (ASR) in order to make our machines more intelligent. While the task and evaluation measure of ASR is clearly agreed by everyone, those of SLU are not so definite. In fact, SLU may be defined in a context of specific applications.

Conventionally, SLU is mainly studied in the context of human-machine dialogue systems which are designed to perform a specific task in a specific domain, such as searching restaurant information or making a flight reservation. Given a task and domain, a set of meanings or concepts can be defined, typically with semantic slots. Then we can easily measure the concept error rate, similarly to the word error rate in ASR, whereas it is often difficult to evaluate the performance of the whole dialogue system as we need to accurately simulate users' behaviors.

Although this traditional view of SLU is still dominant, there has been a significant advancement over a past decade in the coverage of spoken language processing, which has opened new perspectives on SLU. First, the target of spoken dialogue systems has been extended from a simple structured database query to a general information retrieval, including voice search. This causes significant changes in SLU since the "semantic slots" cannot be well defined in the latter task.

Second, the target of ASR has been extended to cover human-human spontaneous speech such as meetings and conversations. Accordingly, SLU needs to be explored for this kind of speech communication when we look for smart archiving or intelligent agents.

In this article, new perspectives on SLU based on these observations are reviewed. In Section II, SLU for the new-generation spoken dialogue systems is discussed. In Section III, SLU for human-human speech communication is addressed, and our new approach focusing on the listener's reaction is introduced.

II. SLU with Information Retrieval (IR)

A. Conventional SLU with definite semantic slots

Conventionally, SLU is formulated to extract meanings from user utterances or their ASR transcripts. The meanings are typically represented by a structure, whose constitutes are made of semantic slots filled by some values, for example, "[destination, Milan]" and "[arrival_date, 12_12_2009]". These are often called concepts, so we can define the concept error rate to measure the SLU performance in a similar way as the word error rate.

There are a number of approaches to extract concepts, either rule-based and statistical-based. A comprehensive review is given by De Mori[1] and by Wang et al[2]. A typical probabilistic formulation to estimate concepts C and a word sequence W for a given speech observation X is defined:

$$p(C,W|X) = p(C)p(W|C)p(X|W)/p(X) \qquad (1)$$

where $p(C)$ is given by a statistical model of concepts such as concept N-gram, $p(W|C)$ by a language model dependent on the concepts, and $p(X|W)$ given by an acoustic model used in ASR.

These approaches assume that a set of concepts is definite given a specific domain and vocabulary. Otherwise, the concept model and the concept error rate cannot be computed. This assumption is consistent with the back-end system which performs a database query based on the SLU result. The semantic slots can be mapped into database fields and values in the relational database (RDB), thus the SLU result is directly translated into an SQL command. In other words, SLU or the

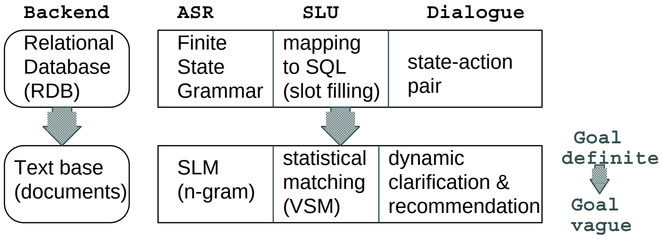
Backend
ASR
SLU
Dialogue
Relational Database (RDB)
Finite State Grammar
mapping to SQL (slot filling)
state-action pair
Text base (documents)
SLM (n-gram)
statistical matching (VSM)
dynamic clarification & recommendation
Goal definite
Goal vague

Fig. 1. From RDB query to IR/QA in dialogue

task itself can be fulfilled by extracting necessary semantic slots, for example, the origin and the destination in train search. This scheme has been successfully applied to a number of applications.

B. From RDB Query to IR

As a majority of data in the world is being stored digitally, but not in a structural way such as RDB, efficient information search/retrieval has become a key technology. Accordingly, speech interfaces for information retrieval (IR), or voice search in general[3], have been investigated. This means that a different methodology is needed in SLU and dialogue management, since the assumption of definite semantic slots does not hold any more.

This paradigm shift is depicted in Fig. 1. In the general document retrieval, neither ASR with a rule-based grammar nor the rule-based SLU is feasible. Even the probabilistic scheme (Eq. 1) is not applicable as long as it assumes a definite set of concepts C. So, SLU is typically based on vector space model (VSM) which consists of word occurrences weighted by a significance measure such as TF-IDF (plus the confidence score of ASR). As for the dialogue strategy, the conventional state-based model is not easily applied since the states are usually pre-defined by the status of semantic slots (filled, not-filled, not-confident,..). Instead, dynamic clarification and disambiguation is required. When the user goal is not definite as in Web surfing, a recommendation function is also desirable.

It is important to note that there are a various types of IR. When the search assumes some structure, for example, directory search has location and listing, we can formulate a noisy channel model[3][4]. On the other hand, when the search space is too large, such as Web search and newspaper article search, there is a little room for SLU in addition to the simple VSM.

Therefore, in this section, we focus on retrieval from document sets in a restricted domain, such as software manuals, cooking recipes, and tourist guides. These can be called knowledge base, since they contain much useful knowledge for problem solving, i.e. trouble shooting, cooking, and tour planning. We presume that this kind of problem solving involves longer interaction than simple search, and thus creates novel features of SLU and dialogue. As we have access to the entire document set, we can model SLU with the matched portion of the documents combined with a sequence of queries.

In the following subsections, two systems we have developed based on this background are introduced. One is retrieval from software support manuals for an automated help desk, and the other is retrieval from tourist guides for a tour guide agent. While the target document is definite in the trouble shooting, there is not a fixed goal in the tour guide task.

C. Interactive Information Retrieval (IR) from Software Support Manuals

Aiming at an automated help desk[5][6], we have designed and implemented voice-interactive system to retrieve from software support manuals. Its detailed description is given in [7].

In this system, a user can make a query stating his trouble. And the system searches for the relevant documents that match the query. In this matching, not only a simple bag of words but also phrases and their dependency relationships are counted[8]. The system clarifies the user queries by asking questions until it narrows down to a small number of documents. The clarification flow is generated on the fly by selecting from a set of questions, which maximizes information gain (IG) or is expected to eliminate the matched documents most effectively[7].

Every time clarification is done, the query is updated and corresponding documents are reduced. The query is in the natural language form constrained by the system. The update of these information states is interpreted as SLU. An example of the dialogue and query update is shown in Fig. 2. Note that the dialogue is dynamically generated without any fixed flow.

The set of clarification questions are prepared beforehand by parsing all the documents[1]. Specifically, we conduct dependency structure analysis for all modifier-head pairs, and calculate the entropy for every head word to measure how much variations of modifiers can appear. For example, the verb “delete” can have various objects such as “application program” and “address book”, but the object of the verb “shutdown” is “computer” in most cases. So, it would be more informative to ask “what did you delete?” than to ask “what did you shutdown?”. These dependency structures almost correspond to case frames, defining a data-driven SLU model.

We have implemented the system with Microsoft software manuals of 40K entries and found that the method achieved a success rate of 83.3% by generating 2.24 questions per dialogue, which is an absolute improvement of 12.6% from the initial query[7].

D. Tour Guide Agent using Question-Answering (QA) Techniques

We have also designed and implemented a virtual tour guide agent based on retrieval from documents on sightseeing spots. Specifically, we have developed a Kyoto navigator using relevant Wikipedia entries and tourist information provided by Kyoto city. Its detailed description is given in [9].

[1] We also enhance with meta-data of knowledge base such as the version of the software.

S1: What is your problem?
U1: I cannot open my file.
S2: What is the file type?
U2: Excel file.
(query: "I cannot open an Excel file")
S3: What is the version of your Excel?
U3: It is version 2002.
(query: "I cannot open an Excel 2002 file")
S4: When did this trouble occur?
U4: When I tried to open it with Explore.
(query: "I cannot open an Excel 2002 file when I try to open with Explore.")
S5: Here are matched documents possibly useful for this trouble.

Fig. 2. Example dialogue and query update by interactive software manual retrieval

While the target (=correct document) in the software support task is uniquely determined as the user has a specific problem to be solved, the dialogue goal of the tour guide is not definite. The user has a vague interest in several spots, and the guide should reply to his requests and also attract potential interests he might have.

To that end, we incorporate the question-answering (QA) technique in addition to information retrieval (IR). The system is able to handle several types of wh-questions such as "person name (who)", "place (where)", "date (when)", and "access information (how)" An off-the-shelf named-entity (NE) tagger is used to label the corresponding entities in the documents. We adopt a standard answer extraction technique, which takes into account (1) the document matching score, (2) the number of matched nouns in the sentence containing the NE, and (3) the number of matched nouns in the phrases having a dependency relationship with the NE.

The QA technique is also applied to generate system-initiative questions to attract the user's interest, which is a kind of information recommendation. This is done by the following steps.

1) Pick up an NE which may attract the user's interest based on the TF-IDF criterion.
2) Substitute the NE with the corresponding interrogative.
3) Delete the subordinate phrases that do not have a dependency relationship with the NE.
4) Transform the sentence into an interrogative form.

An example of the dialogue is shown in Fig. 3. The dialogue is generated automatically from the Wikipedia documents, and the system-initiative utterance S3 corresponds to an information recommendation. In this scheme, SLU is realized by the QA module implicitly. In the voice-interactive systems, one of the difficulties is in the topic and focus detection, since many questions contain ellipsis and anaphora which should be resolved, for example, "When was (it) built?". Moreover, the system-initiative recommendation would be irrelevant if it is out of focus of the current dialogue.

A simple heuristics on the metadata of the documents (=sightseeing spot) is adopted in this system, although topic detection can be generally formulated as a classification problem[10],

U1: Please explain Golden Pavilion.
S1: Golden Pavilion is one of the buildings in the Rokuon-ji in Kyoto, and is the main attraction of the temple sites. The entire pavilion except the basement floor is covered with a pure gold leaf.
U2: When was it built?
S2: Golden Pavilion was originally built in 1397 to serve as a retirement villa for Shogun Ashikaga Yoshimitsu.
..(silence)..
S3: Well then, do you know what was awarded to this temple in 1994?
U3: No, please tell me.
S4: It was awarded as listing on the UNESCO World Heritage in 1994.
U4: How can I get there?
.....

Fig. 3. Example dialogue in the Kyoto navigator

III. SLU for Human-Human Speech Communication

SLU is now being explored for human-human speech communication, which includes monologue addressed to audiences such as lectures, dialogues, and multi-party conversations such as meetings.

A. Rich Transcription and Summarization

Automatic transcription of spontaneous speech such as lectures and meetings essentially should involve SLU since faithful transcription is not necessarily useful because of the existence of disfluencies and the lack of sentence and paragraph markers. Therefore, a number of research projects, including those sponsored by DARPA and NIST, are oriented to "rich transcription", which involves annotation of end-of-sentence markers and disfluency phenomena[11] [12][13]. In the Corpus of Spontaneous Japanese (CSJ)[14], too, these kinds of annotations were attached and extensive works have been conducted[15].

As a whole, these tasks are addressed with machine learning approaches; a set of relevant features, including lexical, prosodic and even syntactic features[16], are counted. And statistical classifiers such as SVM (Support Vector Machines) and CRF (Conditional Random Fields) are trained.

In addition to disfluency deletion and punctuation insertion, there are a number of edits to be done for the final clean transcripts, including correction of colloquial expressions and dropped words. Handling these phenomena is particularly important in formal settings such as public speeches and congressional meetings. We have been developing an automatic transcription system for the Japanese National Diet (=Congress)[17], and implementing the post-processing process with WFST (Weighted Finite State Transducers)[18].

These processes of rich transcription are important, but they are essentially shallow processing rather than understanding [2].

Dialogue act tagging[19], which determines the intent type of user utterances, leads to a high-level SLU. Based on it, action item identification is also explored[20].

Information extraction (IE)[21][22], which identifies named entities (NEs) and their relationships, is regarded as a form of document understanding. It is often formulated in a restricted domain, for example, merger-acquisition relationship in economy and industry-related articles. A new direction of SLU may be in the combination of the information extraction and dialogue act tagging.

Speech summarization is to extract important portions and generate compact outputs. SLU is apparently involved in speech summarization, at least by human. One of the most popular automatic summarization methods is based on Maximal Marginal Relevance (MMR)[23][24], which is defined as the similarity in a vector space model (VSM), often reduced by LSA, and extracts sentences which best match the entire document and yet are different from each other. The approach is similar to the incremental search strategy discussed in Section II-D in that the VSM is used for SLU.

B. Key Spot Detection based on Listener's Reaction

The underlying idea of the conventional speech summarization including other methods[25] relies on the features, such as lexical, prosodic and discourse features[26], of the main speaker's speech, which are presumably related to the core or emphasized portion of the speech. This approach is typically called "content-based" indexing, because it requires processing, such as ASR and lexical analysis of the audio content to be indexed.

As opposed to the conventional content-based approach, we propose a novel approach to extraction of important portions from speech, focusing on the audience's reaction. Even if we do not understand the speech given in a foreign language, we can easily see which part is funny and which part is important by observing the audience's reaction such as laughing and nodding.

For a multi-modal study of speech communication, we have started a new project on multi-modal recording and analysis of poster presentations[27]. Typically in poster sessions, a presenter explains his work to a small audience using a poster, and the audience gives feedback in real time by nodding or backchannels, and occasionally makes questions and comments. Thus, the poster session has a mixture of characteristics of lectures and meetings: There is a main speaker, but anyone can be highly involved in the conversation at a certain point. We expect that the audience's reaction is more apparent in poster sessions than in oral presentations, because the size of audience is smaller and the style is more interactive.

As a first step of the "interaction-based" indexing, we focus on the audience's backchannel responses, rather than investigating overall prosodic patterns as in adopted in former studies on "hot-spot" detection[28][29].

By backchannel responses (*Aizuchi* in Japanese), we mean the listener's verbal short response, which expresses his state of the mind during the conversation. Its prototypical lexical entries include "*hai*" in Japanese and "yeah" or "okay" in English. Note that many of them are non-lexical and used only for backchannel responses, such as "*hu:n*" in Japanese and "uh-huh" in English. It is well-known that the backchannel response suggests that the listener is understanding what is being said. Moreover, we hypothesize that the audience signals his interest level with the syllabic and prosodic pattern of the backchannel responses. Thus, we expect that this information will be useful for identifying "key spots" in the presentation and indexing for efficient access to the archive.

One of the problems in this task is to distinguish the patterns of reactive tokens which are related with the interest level. For that purpose, we investigated the frequency of reactive tokens used by the primary listener and by the general audience. By "primary listener", we mean the audience who had raised a question to the presenter and listened to the reply during a segment of the poster session. We found a significant difference in the frequency of "*hai* (yes/yeah)" between the two conditions. Since the main function of "*hai*" is presumably acknowledgment, suggesting "I hear/understand you", it is reasonable that the person who raised question should have the courtesy to acknowledge the answer.

So we focus on three reactive tokens of "*hu:N*", "*he:*" and "*a:*", which are not used for acknowledgment, and found that they are correlated with the interest level when uttered in some specific prosodic patterns. It is observed that prolonged "*hu:N*" means interest and surprise while "*a:*" with higher pitch or larger power means interest. On the other hand, "*he:*" can be emphasized in all three prosodic features (duration, power, pitch) to express interest and surprise.

We have also investigated a method to extract these reactive tokens robustly in multi-party conversations[30], and designed a graphical user interface to browse conversations through these tokens.

There are more multi-modal features which are related with the listener's reaction and may be useful for identifying key spots of speech communication. This is an indirect approach of SLU in that the system does not understand the main speaker's utterances, but watches the reactions of the audience, who has an apparently better understanding ability. The approach can also be complemented with the orthodox content-based SLU.

IV. Conclusion

In this article, two new perspectives on SLU have been reviewed. One is based on the paradigm shift of spoken language systems from a simple database query to a general information retrieval. This shift has changed SLU from semantic slot extraction to a vector space model. Although it is doubtful that this model can be called "understanding", we still note an importance of understanding some structures such

[2] Congressional stenographers say they are not sure they "understand" the speech during shorthand transcription, but "do not hear" disfluencies.

as dependency relationships and discourse structures in many dialogue applications.

The other is exploration of SLU in human-human communication. In this field, “understanding” in a strict sense is very difficult, and we introduce another approach which focuses on the human understanding process. This approach is still preliminary and should be investigated more and also combined with the conventional SLU methods.

Acknowledgment

A part of the works presented in this article is indebted to Dr. Teruhisa Misu and Mr. Katsuya Takanashi. The latter part of the work is now funded by JST, CREST.

References

[1] R.De Mori. Spoken language understanding: a survey. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[2] Y.-Y.Wang, L.Deng, and A.Acero. Spoken language understanding. *Signal Processing Magazine*, 22(5):16–31, 2005.

[3] Y.-Y.Wang, D.Yu, Y.-C.Ju, and A.Acero. An introduction to voice search. *Signal Processing Magazine*, 25(3):29–38, 2008.

[4] Y-I.Song, Y-Y.Wang, Y-C.Ju, M.Seltzer, I.Tashev, and Al.Acero. Voice search of structured media data. In *Proc. IEEE-ICASSP*, pages 3941–3944, 2009.

[5] D.Griol, G.Riccardi, and E.Sanchis. A statistical dialog manager for the LUNA project. In *Proc. INTERSPEECH*, pages 272–275, 2009.

[6] D.Suendermann, K.Evanini, J.Liscombe, P.Hunter, K.Dayanidhi, and R.Pieraccini. From rule-based to statistical grammars: Continuous improvement of large-scale spoken dialog systems. In *Proc. IEEE-ICASSP*, pages 4713–4716, 2009.

[7] T.Misu and T.Kawahara. Dialogue strategy to clarify user’s queries for document retrieval system with speech interface. *Speech Communication*, 48(9):1137–1150, 2006.

[8] Y.Kiyota, S.Kurohashi, T.Misu, K.Komatani, T.Kawahara, and F.Kido. Dialog navigator: A spoken dialog Q-A system based on large text knowledge base. In *Proc. ACL-03*, volume Interactive Poster & Demo., pages 149–152, 2003.

[9] T.Misu and T.Kawahara. Speech-based interactive information guidance system using question-answering technique. In *Proc. IEEE-ICASSP*, volume 4, pages 145–148, 2007.

[10] I.R.Lane, T.Kawahara, T.Matsui, and S.Nakamura. Out-of-domain utterance detection using classification confidences of multiple topics. *IEEE Trans. Audio, Speech & Language Process.*, 15(1):150–161, 2007.

[11] J.S.Garofol, C.D.Laprun, and J.G.Fiscus. The RT-04 spring meeting recognition evaluation. In *NIST Meeting Recognition Workshop*, 2004.

[12] Y.Liu, E.Shriberg, A.Stolcke, B.Peskin, J.Ang, D.Hillard, M.Ostendorf, M.Tomalin, P.Woodland, and M.Harper. Structural metadata research in the EARS program. In *Proc. IEEE-ICASSP*, volume 5, pages 957–960, 2005.

[13] Y.Liu, E.Shriberg, A.Stolcke, D.Hillard, M.Ostendorf, and M.Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Trans. Audio, Speech & Language Process.*, 14(5):1526–1540, 2006.

[14] K.Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, 2003.

[15] Sadaoki Furui and Tatsuya Kawahara. Transcription and distillation of spontaneous speech. In J.Benesty, M.M.Sondhi, and Y.Huang, editors, *Springer Handbook on Speech Processing and Speech Communication*, pages 627–651. Springer, 2008.

[16] T.Kawahara, M.Saikou, and K.Takanashi. Automatic detection of sentence and clause units using local syntactic dependency. In *Proc. IEEE-ICASSP*, volume 4, pages 125–128, 2007.

[17] Y.Akita, M.Mimura, and T.Kawahara. Automatic transcription system for meetings of the Japanese. In *Proc. INTERSPEECH*, pages 84–87, 2009.

[18] G.Neubig, S.Mori, and T.Kawahara. A WFST-based log-linear framework for speaking-style transformation. In *Proc. INTERSPEECH*, pages 1495–1498, 2009.

[19] E.Shriberg, R.Dhillon, S.Bhagat, J.Ang, and H.Carvey. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proc. SIGDial*, pages 97–100, 2004.

[20] F.Yang, G.Tur, and E.Shriberg. Exploiting dialog act tagging and prosodic information for action item identification. In *Proc. IEEE-ICASSP*, pages 4941–4944, 2008.

[21] L.Ramshaw and R.M.Weischedel. Information extraction. In *Proc. IEEE-ICASSP*, volume 5, pages 969–972, 2005.

[22] R.Grishman. Discovery methods for information extraction. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 243–247, 2003.

[23] J.Carbonell and J.Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. ACM SIG-IR*, 1998.

[24] Y.Gong and X.Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proc. SIG-IR*, 2001.

[25] S.Furui. Recent advances in automatic speech summarization. In *Proc. IEEE/ACL Workshop Spoken Language Technology*, 2006.

[26] T.Kawahara, M.Hasegawa, K.Shitaoka, T.Kitade, and H.Nanjo. Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers. *IEEE Trans. Speech & Audio Process.*, 12(4):409–419, 2004.

[27] T.Kawahara, H.Setoguchi, K.Takanashi, K.Ishizuka, and S.Araki. Multimodal recording, analysis and indexing of poster sessions. In *Proc. INTERSPEECH*, pages 1622–1625, 2008.

[28] B.Wrede and E.Shriberg. Spotting “hot spots” in meetings: Human judgments and prosodic cues. In *Proc. EUROSPEECH*, pages 2805–2808, 2003.

[29] D.Gatica-Perez, I.McCowan, D.Zhang, and S.Bengio. Detecting group interest-level in meetings. In *Proc. IEEE-ICASSP*, volume 1, pages 489–492, 2005.

[30] K.Sumi, T.Kawahara, J.Ogata, and M.Goto. Acoustic event detection for spotting hot spots in podcasts. In *Proc. INTERSPEECH*, pages 1143–1146, 2009.