

Editorial

Introduction to the Special Issue on Spoken Language Understanding in Conversational Systems

1. Introduction and previous work

Understanding spoken language is about extracting the *meaning* from speech utterances. Although there continues to be endless debates in linguistics, philosophy, psychology, and neuroscience on what constitutes the meaning of a natural language utterance (Jackendoff, 2002), for the purpose of human–computer interactive systems, “meaning” is regarded as a representation that can be executed by an interpreter in order to change the state of the system. In such systems, understanding spoken language involves automatic speech recognition (ASR) and spoken language understanding (SLU)—a transduction of the recognition result to an interpretable representation.¹

There are a number of aspects to SLU that makes this a challenging task. The most challenging of them all is the issue of meaning representation. Human language expresses meaning through a variety of surface forms, for example, prosody, lexical choice, syntax. The same meaning can be expressed in many different surface forms and also the same surface form can express many different meanings. These aspects are further accentuated in conversational systems where the dialog context plays a significant role in the meaning of an utterance. Designing a representation that captures this rich expressivity in full generality is a daunting task. So, in order to build practical systems, meaning representations tend to be crafted based on the desired capabilities

of an application. However, such representations constrain the portability of the system to new domains.

Another aspect that is particularly acute for spoken language understanding is robustness to noise in the input. Unlike text understanding, the input to SLU is noisy since it is the output of a speech recognizer. In addition to this noise, spoken language is rampant with disfluencies such as filled pauses, false starts, repairs and edits. Hence, in order to be robust, the SLU architecture needs to cater and cope with the noisy input from the beginning and not as an after thought. Also, the meaning representation should be such that it can support robust inference even in the presence of noise.

We see many of these aspects addressed in this Special Issue with innovative solutions to each of these problems. Before summarizing the papers in this collection and discussing their contributions to SLU, we first begin by reviewing the history of spoken language understanding systems and cataloging them in terms of the increasing sophistication of the meaning representations. Finally, we will provide some indications to future directions to the SLU research and its role beyond conversational systems.

2. History of SLU

From the early 1990s, there have been a variety of practical goal-oriented spoken dialog systems (SDSs) for applications in limited domains. These systems, typically, identified the users’ intents expressed in natural language, and acted on them appropriately, in order to satisfy the users’ requests.

¹ Although, ASR and SLU are usually regarded as two stages of processing, there has been work on tightly integrating them (Hakkani-Tür et al., in press).

The intents and actions were predefined to suit the capabilities of the system. In such systems, typically, the speaker's utterance is first recognized using an automatic speech recognizer (ASR). Then, the intent of the speaker is identified from the recognized sequence, using a spoken language understanding (SLU) component. Finally, the role of the dialog manager (DM) is to interact with the user in a natural way and help the user to achieve the task that the system is designed to support.

The first generation SDSs were typically machine-initiative systems, where the users answered the computer prompts using a set of permitted phrases (Wilpon et al., 1990) (e.g. *Please say hotel reservation or car reservation*). In such a directed dialog, "understanding" is reduced to the task of detecting one of the keywords allowed in the users' utterances.

In the second generation, SDSs employed a more sophisticated understanding system which allowed for the users to talk naturally about a *single* task. Such systems were first developed under government-funded projects. In the early 1990s, DARPA initiated the Airline Travel Information System (ATIS) project, with the aim of integrating the efforts of the speech and language processing communities toward sophisticated dialog systems (Price, 1990). In this task, the users could inquire about flight information, for example, *I want to fly to Boston from New York next week, Does this flight serve meals*. The meaning representation in such systems were frames for each sub-task (e.g. *booking, inquiry*) and each frame had predefined slots that were appropriate for that sub-task. Understanding in these systems amounted to extracting specific fillers for each slot (e.g. *Destination* and *Departure Date*) in a given frame. To fill the slots of a frame, prototypical systems either employed a data-driven statistical approach (mostly from the speech processing community), such as AT&T's CHRONUS (Pieraccini et al., 1992) and BBN's Hidden Understanding Models (Miller et al., 1994) or a knowledge-based approach (mostly from the computational linguistics community), such as MIT's TINA (Seneff, 1992), CMU's Phoenix (Ward and S.Issar, 1994), and SRI's Gemini (Dowding et al., 1993).

The third generation of SDSs focused more on the dialog component, such as the ESPRIT Sundial project (Peckham, 1991) and the DARPA Communicator programme (Walker et al., 2002). For example, the Communicator programme aimed at rapid development of speech-enabled dialog systems with

advanced conversational capabilities. Eight systems from AT&T, BBN, University of Colorado, Carnegie Mellon University, IBM, Lucent Bell Labs, MIT, and SRI participated in the 2001 evaluation. These systems supported complex dialogs with the users to arrange their travel plans. The goal of the travel system was to provide a wide range of travel related services including multi-leg flights, hotel and car arrangements. Such programmes have left a test-bed for further SLU research and have led to similar mixed- or machine-initiative commercial systems and research projects (den Os et al., 1999; Strik et al., 1997; FASiL, 2002, among others.). Here, it is also worth mentioning the multi-modal and multi-media dialog systems and projects such as the ESPRIT MASK (Lamel et al., 1998) and AT&T MATCH (Johnston et al., 2002).

A related but independent effort was the Message Understanding Conferences, where the aim is to extract information from natural language texts (MUC-7, 1998). Similar to the ATIS project, participating sites have built systems filling templates of named entities (NEs). A named entity is defined to be either a proper name, i.e., location, person, or organization name, or a numeric entity, i.e., date, time, amount, or percentage. For the example utterance above, both *Boston* and *New York* would be the named entities of type *Location*, and *next week* would be a named entity of type *Date*. Note that defining departure and arrival locations and dates as separate named entities, the overall ATIS task can be seen as a named entity extraction problem, however from spoken language utterances.

A different approach to SLU is to both determine the intent of the user and extract the corresponding arguments as in the AT&T *How May I Help You?*SM (HMIHY) spoken dialog system used for customer care centers (Gorin et al., 1997). The users of the system are greeted by the open-ended prompt *How May I Help You?* encouraging them to talk naturally. The system identifies the customer's intent (call-type), by classifying the call into one of a few application-specific categories and extract relevant information for that category (or named entities). As an example, consider the utterance *I have a question about my June bill*. Assuming that the utterance is recognized correctly, the corresponding intent (call-type) would be *Ask(Bill)* and the argument for this call-type, i.e., the named entity *Date* would be extracted as *June*. Then the action that needs to be taken depends on the DM. It may ask the user to further specify the problem or

route this call to the billing department. Following the HMIHY system, a number of similar systems have been built (Chu-Carroll and Carpenter, 1999; Natarajan et al., 2002; Gupta et al., in press, among others).

In all these previous approaches, the representation for the utterance meaning has heavily relied on the corresponding application. The representation is carefully crafted for optimum performance. For example in ATIS, the slots are related to flight reservations, such as arrival and departure cities. In HMIHY, the meaning representation is the call-type and the corresponding arguments (named entities). Note that in the AI community there have been many related though independent natural language understanding systems such as TRAINS (Allen et al., 1995), and research projects such as the DAR-PA PAL Programme (PAL, 2003).

3. About this volume

In the past few years, there has been a substantial increase in interest in information extraction from the NLP community, question-answering in the information retrieval community, and spoken dialog systems in the speech processing community. Spoken language understanding is an especially attractive topic for cross-fertilization of ideas between speech, AI, IR and NLP communities. This Special Issue is in part a compilation of extended versions of the papers presented at the Spoken Language Understanding Workshop organized at NAACL 2004 in Boston. This volume presents a balanced mix of papers on SLU from both academic and industrial research laboratories.

The articles in this Special Issue address a range of topics concerning SLU—robust inference using machine learning techniques, tightly coupling SLU with the speech recognizer in order to minimize understanding error rate, bootstrapping a portable meaning representation, rapid prototyping of SLU-based applications, and the use of prosodic, discourse and pragmatic features for improving SLU. There are also papers that describe SLU as a component technology and discuss their role in end-to-end applications.

3.1. Data-driven approaches to SLU

In this collection, most SLU systems are data driven. With the availability of data from deployed

dialog systems, this direction of research holds the greatest promise. The motivation for this approach is robustness to noise and rapid prototyping of SLU.

In his paper, Haffner presents a variety of machine learning techniques for intent determination, and compares them based on a number of criteria, such as performance, robustness, and CPU and memory requirements.

He and Young address the issue of robustness to noise, and portability and extensibility of a representation called hidden vector state (HVS) model for SLU and contextualize this model for the ATIS domain. In the HVS model, the context is encoded as a stack-oriented state vector, enabling the model to capture the hierarchical semantic tree structure.

In order to improve the understanding error rate, Saraclar and Roark, in their paper, show that the understanding error rate can be significantly reduced by discriminatively training the language models using the SLU performance as the optimization criterion. This paper can also be seen as a tighter coupling of the language model with the SLU.

3.2. Tighter coupling of ASR and SLU

As mentioned earlier, the input to SLU is noisy and using the one-best output of the ASR for understanding is bound to be suboptimal. In order to alleviate the errors in the one-best output, SLU could consider multiple hypotheses from ASR represented as a word lattice. This theme is reflected in the following three papers. These papers structure the SLU as weighted finite state transducers (FSTs) which allows them to exploit the word lattice output by the recognizer efficiently.

Raymond et al. use a two-level knowledge representation: the first language dependent level is made of a set of conceptual constituents. The second language independent level consists of semantic relations between these conceptual constituents. They use FSTs to represent and extract conceptual constituents from the lattice output of an ASR. Each hypothesis is scored using the path posterior probabilities in the lattice. The semantic relations are found using semantic composition rules.

Wutiwiwatchai and Furui use a three-stage approach, where first the concepts are extracted from a given word sequence, then the goals are identified, and concept values are determined. In their system, concept extraction also uses weighted FSTs. They

use a new logical N-gram model for concept extraction and concept value estimation. For goal identification, they use a classification-based approach, and compare artificial neural networks, support vector machines and MAP-based methods.

If we consider SLU to be a transduction of the speech signal to a symbol sequence representing the meaning of the utterance, machine translation—transducing a source language speech signal to a target language symbol sequence is an extreme example of such a transduction. The paper by Ruiqiang Zhang et al. on integration of speech recognition and machine translation presents such an approach for speech-to-speech translation. The paper focuses on tight integration of speech recognition and machine translation in order to mutually compensate for errors in each component, by translating the word lattice output by the speech recognizer instead of translating the first-best speech recognition output. The approach is used to translate Japanese-to-English from a tourist travel domain.

3.3. SLU for end-to-end systems

The next set of papers address the issues involved in designing SLU for end-to-end systems for applications. They address issues pertaining to crafting the meaning representation for the application and rapidly bootstrapping an SLU by combining data and hand-crafted application-specific grammars. A particularly acute problem for data-driven SLU techniques is one of data acquisition and annotation. Several approaches are presented in this set of papers that alleviate this problem.

The paper by Boye et al. discusses the application of robust spoken language understanding in the context of a computer game. The paper focuses on the semantic representation used in the system and the robust inference mechanisms associated with this semantic representation. The semantic representation involves typed entities, predicate-argument information and function abstraction to account for underspecification. Robust inference is performed as constraint solving using type resolution and function application, in contrast to the machine learning techniques for classification and slot-filler extraction. The system accuracy is measured in terms of concept accuracy on the resulting semantic representation.

Hardy et al. present the AMITIÉS spoken dialog system, a natural language customer service application system for a phone banking call center. For

understanding the user utterances, they first determine the intent and the dialog act, and extract named entities and slot-filling frame agents using probabilistic methods.

Huang and Cox tackle the problem of training data preparation, which is a labor-intensive and time-consuming task. They present several techniques using a phone recognizer to build call routing systems from untranscribed user utterances. They learn salient phone phrases to identify the user intents.

Wang and Acero propose a knowledge-assisted example-based statistical model for SLU. They present a grammar authoring tool, called SGStudio, where the application developers form some typical example sentences from which context free grammars are learned.

3.4. Exploiting linguistic constraints for SLU

As is well known, the meaning of an utterance is only partly conveyed by the words and syntax of an utterance. For speech, the prosody of an utterance carries a great deal of information that contributes to the meaning of the utterance. Furthermore, in a conversational system the effects of the dialog history and the pragmatic salience of the utterances contribute significantly towards the meaning of an utterance. Although this topic is still in its infancy, there is a growing interest in exploiting such information for improving SLU as reflected by the following two papers in this collection.

Higashinaka et al. show that discourse related features can be combined with acoustic and language model related features to improve confidence scoring of concept values in a spoken dialog system for weather information service domain.

Tong Zhang et al. propose and describe the extraction of two linguistic concepts *contrast* and *focus* for the extraction of pragmatic and semantic salience from spontaneous user utterances in the framework of a tutoring application. They use word similarity and dissimilarity analysis, part-of-speech tags, and prosodic features such as duration and pitch related features for the extraction.

4. Future directions

The papers in this issue echo similar themes in terms of the future directions for research in spoken language understanding:

4.1. Semantic representation

The issue of representing the meaning of an utterance in a form that is conducive for robust inference and portable across domains and yet exploits the constraints of the application is still the single most challenging problem of spoken language understanding. This is partly due to the fact that the semantic representation needs to cater to different kinds of spoken dialog applications ranging from Call Routing, Information Access, Transaction-oriented, Problem Solving, Tutoring, Games/Entertainment.

4.2. Robust inference

The representation also has a significant impact on the interface between the ASR–SLU and SLU–DM. In order to support robust inference, these interfaces need to represent and process uncertain input and output such as the lattices of recognized utterances and meaning expressions with confidence values.

A promising direction for robust SLU is the use of machine learning techniques. However, the bottleneck for these techniques is that they rely heavily on data that is hard to collect and tedious to transcribe and annotate. There is a growing need for techniques that balance data-driven techniques with knowledge-based techniques in a unified way to address the issue of rapid bootstrapping and robust inference in SLU.

4.3. Availability of real-world data

Many of the speech and NLP tasks (speech recognition, translation, part-of-speech tagging, named entity extraction, semantic role labeling) have benefited significantly, thanks to the involvement of a community of researchers in benchmarking algorithms and comparing results on shared data sets. Such a community-wide effort would not have been possible if not for the availability of shared, annotated datasets. SLU research on the other hand has been studied in isolated groups with specialized semantic representations and idiosyncratic/proprietary data sets. This makes comparison of different SLU techniques extremely difficult, and results in duplication of effort. In order for SLU to reap the benefits of a unified effort, it is imperative to create and share an annotated dialog corpus from a real-world application.

The collection of dialog data has a set of challenges of its own. An important issue concerns with the nature of the data itself. Should the data be from human–human conversations (since we eventually would like to emulate a human as a conversational partner), or from human–machine dialog systems (since humans display difference in styles when conversing with machines and other humans)? An intermediate approach is to collect wizard data. However, this approach has the limitation of being constrained to a few scenarios and the methodology is not scalable. Another issue is the confidentiality of the data, which prevents the sharing of some datasets.

4.4. Evaluation

Another issue that prevents the comparison of the performances of various systems is the lack of community-wide accepted evaluation metrics. SLU can be evaluated at the utterance (i.e., query accuracy, as in ATIS or intent determination accuracy, as in HMIHY) or at the dialog level (i.e., PARADISE (Walker et al., 1997), as in Communicator). Even though there have been many proposals for evaluation, there is not a single method that is used by all approaches.

To conclude, we believe that SLU research can benefit significantly from funding for data collection and emphasis on evaluation metrics that can be used for cross-system comparisons.

Acknowledgments

We would like to thank all the authors for contributing to this volume, Speech Communication Journal Editor-in-Chief Renato De Mori, and Elsevier Editorial Production Department (Linda Mulder and Mary Lynn van Dijk) for their help and patience, and the anonymous reviewers for selecting and improving the presentation of the papers.

References

- Allen, J.F., Schubert, L.K., Ferguson, G., Heeman, P., Hwang, T., Kato, T., Light, M., Martin, N.G., Miller, B.W., Poesio, D.R., Traum, D.R., 1995. The TRAINS project: a case study in building a conversational planning agent. *J. Exp. Theor. AI (JETAI)* 7, 7–48.
- Chu-Carroll, J., Carpenter, B., 1999. Vector-based natural language call routing. *Comput. Linguist.* 25 (3), 361–388.
- den Os, E., Boves, L., Lamel, L., Baggia, P., 1999. Overview of the ARISE project. In: *Proc. Sixth European Conf. in Speech*

- Communication and Technology (EUROSPEECH). Budapest, Hungary, pp. 1527–1530.
- Dowding, J., Gawron, J.M., Appelt, D., Bear, J., Cherny, L., Moore, R., Moran, D., March 1993. Gemini: a natural language system for spoken language understanding. In: Proc. ARPA Workshop on Human Language Technology, Princeton, NJ.
- FASiL, 2002. FASiL: flexible and adaptive spoken language and multi-modal interfaces. Available from: <<http://www.fasil.co.uk>>.
- Gorin, A.L., Riccardi, G., Wright, J.H., 1997. How May I Help You? *Speech Commun.* 23, 113–127.
- Gupta, N., Tur, G., Hakkani-Tür, D., Bangalore, S., Riccardi, G., Rahim, M., in press. The AT&T spoken language understanding system. *IEEE Trans. Speech Audio Process.*
- Hakkani-Tür, D., Béchet, F., Riccardi, G., Tur, G., in press. Beyond ASR 1-Best: using word confusion networks in spoken language understanding. *Comput. Speech Language.*
- Jackendoff, R., 2002. *Foundations of Language*. Oxford University Press (Chapter 9).
- Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., Maloor, P., 2002. MATCH: An architecture for multimodal dialogue systems. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA.
- Lamel, L., Bennacef, S., Gauvain, J., Dartigues, H., Temem, J., 1998. User Evaluation of the MASK Kiosk. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP), Sydney, Australia.
- Miller, S., Bobrow, R., Ingria, R., Schwartz, R., 1994. Hidden understanding models of natural language. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL), Las Cruces, NM.
- MUC-7, 1998. Proc. 7th Message Understanding Conference (MUC-7), Fairfax, VA.
- Natarajan, P., Prasad, R., Suhm, B., McCarthy, D., 2002. Speech enabled natural language call routing: BBN call director. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP), Denver, CO.
- PAL, 2003. DARPA personalized assistant that learns (PAL) programme. Available from: <<http://www.darpa.mil/ipto/programs/pal>>.
- Peckham, J., 1991. Speech understanding and dialogue over the telephone, an overview of the ESPRIT SUNDIAL project. In: Proc. DARPA Speech and Natural Language Workshop, Pacific Grove, CA.
- Pieraccini, R., Tzoukermann, E., Gorelov, Z., Gauvain, J.-L., Levin, E., Lee, C.-H., Wilpon, J.G., 1992. A speech understanding system based on statistical representation of semantics. In: Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), San Francisco, CA.
- Price, P.J., 1990. Evaluation of spoken language systems: the ATIS domain. In: Proc. DARPA Workshop on Speech and Natural Language, Hidden Valley, PA.
- Seneff, S., 1992. TINA: a natural language system for spoken language applications. *Comput. Linguist.* 18 (1), 61–86.
- Strik, H., Russel, A., van den Heuvel, H., Cucchiari, C., Boves, 1997. A spoken dialogue system for the Dutch public transport information service. *Internat. J. Speech Technol.* 2 (2), 119–129.
- Walker, M., Rudnicky, A., Prasad, R., Aberdeen, J., Bratt, E., Garofolo, J., Hastie, H., Le, A., Pellom, B., Potamianos, A., Passonneau, R., Roukos, S., Sanders, G., Seneff, S., Stallard, D., 2002. Darpa communicator: cross-system results for the 2001. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP), Denver, CO.
- Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A., 1997. PARADISE: a framework for evaluating spoken dialogue agents. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL)—Conference of the European Chapter of the Association for Computational Linguistics (EACL), Madrid, Spain.
- Ward, W., Issar, S., 1994. Recent improvements in the CMU spoken language understanding system. In: Proc. ARPA Human Language Technology Conference (HLT) Workshop. pp. 213–216.
- Wilpon, J.G., Rabiner, L.R., Lee, C.-H., Goldman, E.R., 1990. Automatic recognition of keywords in unconstrained speech using hidden markov models. *IEEE Trans. Speech Audio Process.* 38 (11), 1870–1878.

Srinivas Bangalore
 Dilek Hakkani-Tür
 Gokhan Tur
 AT&T Labs—Research
 Florham Park
 NJ 07932
 USA

E-mail addresses: srini@research.att.com (Srinivas Bangalore)
dtur@research.att.com (Dilek Hakkani-Tür)
gtur@research.att.com (Gokhan Tur)