© IMAGESTATE

Renato De Mori, Frederic Béchet, Dilek Hakkani-Tür, Michael McTear, Giuseppe Riccardi, and Gokhan Tur

Spoken Language Understanding

[Interpreting the signs given by a speech signal]

Semantics deals with the organization of meanings and the relations between sensory signs or symbols and what they denote or mean [29]. Computational semantics performs a conceptualization of the world using computational processes for composing a meaning representation structure from available signs and their features present, for example, in words and sentences. Spoken language understanding (SLU) is the interpretation of signs conveyed by a speech signal. SLU and natural language understanding (NLU) share the goal of obtaining a conceptual representation of natural language sentences. Specific to SLU is the fact that signs to be used for interpretation are coded into signals along with other information such as speaker identity. Furthermore, spoken sentences often do not follow the grammar of a language; they exhibit self-corrections, hesitations, repetitions, and other irregular phenomena. SLU systems contain an automatic speech recognition (ASR) component and must be robust to noise due to the spontaneous nature of spoken language and the errors introduced by ASR. Moreover, ASR components output a stream of words with no structure information like punctuation and sentence boundaries. Therefore, SLU systems cannot rely on such markers and must perform text segmentation and understanding at the same time.

Obtaining meaning from speech is a complex process and many different approaches and models have been proposed. Systems developed in the 1970s and the 1980s mostly performed syntactic analysis on the best sequence of words hypothesized by an ASR system and used nonprobabilistic rules for mapping syntactic structures into semantic ones expressed as logic formulas. An interesting discussion on computer structures for semantic representations considered in this period can be found in [29]. Meaning representation and approaches for obtaining these representations from words are discussed in this article. Basic related problems are reviewed in [15]. In the 1990s, the need emerged for testing SLU processes on large corpora that could also be used for automatically estimating some model parameters. Probabilistic finite-state interpretation models and grammars were also introduced for dealing with ambiguities introduced by model imprecision. Systems based on these approaches are discussed in this article and are also reviewed in chapter 14 of [6].

Some other approaches transform signals directly into basic semantic constituents to be further composed into semantic structures. This integration of the ASR and SLU processes, which is discussed in this article, generates multiple SLU hypotheses to be further validated using constraints imposed by the context in which a sentence is interpreted.

Digital Object Identifier 10.1109/MSP.2008.918413

1053-5888/08/$25.00©2008IEEE

The level of complexity needed in order to represent the meaning of a spoken utterance depends mainly on the application targeted. There are three main application domains for SLU: spoken dialog systems, speech information retrieval (or speech mining), and speech translation.

We will not address speech translation in this article as the SLU models needed are heavily dependent on the translation method used, and this is out of the scope of this article. Speech mining applications are mostly focused on the retrieval of semantic information such as entities (named entities or application-dependent concepts), themes, and opinions. For this purpose, most of the time a flat semantic representation, like an attribute/value sequence, is used to represent the interpretation of a spoken utterance. Spoken dialog systems need advanced SLU models in order to implement dialog applications that go beyond call routing or form filling. For example, the European project LUNA (www.ist-luna.eu) defines three levels of complexity for dialog applications. The first level includes the translation process from words into basic conceptual constituents (generation of semantic concepts). This level of detail is sufficient for applications such as call routing and utterance classification with a mapping to disjoint categories. The second level performs semantic composition on basic constituents for applications like call routing with utterance characterization (finer-grain comprehension), question/answering, and inquiry qualification. At the third level, a broad context is taken into account for context-sensitive validation in complex spoken dialog applications and inquiry qualification considering an utterance as a set of sub-utterances and the interpretation of one sub-utterance being context sensitive to the others. The different semantic models and interpretation processes to be presented in this article will all be focused on dialog applications belonging to one of these levels.

Applications are effective if the systems have self-diagnosis capabilities in order to commit transactions or perform other actions. Confidence in SLU will also be discussed. Speech data that are not interpreted with high confidence can be proposed for manual annotation and used for successive model refinement. Active learning for this purpose is also discussed in this article.

COMPUTER REPRESENTATIONS OF MEANING

A meaning representation language (MRL) has its own syntax and semantics and should follow a representation model coherent with a semantic theory, taking into account intension and extension relations, reasoning, composition of semantic constituents into structures, and procedures for relating them to signs. Designing a meaning representation that can capture the rich expressivity of spoken language is difficult. Therefore, in order to build practical systems, meaning representations tend to be crafted based on the desired capabilities of each application.

The semantic knowledge of an application is stored in a knowledge base. A convenient approach to reasoning about semantic knowledge is to represent it as a set of logic formulas. Formulas contain variables that are bound by constants and may be typed. First-order or higher-order logics can be used. Concepts carried by signs are asserted by an interpretation process. New assertions can be obtained from asserted concepts by an inference process.

Task-dependent representations constrain the portability of a system to new domains and applications. Recently, there have been two new semantic representation frameworks that are widely accepted: FrameNet and PropBank.

Computer semantics has to be based on models that represent knowledge with schemata including procedures for hypothesizing semantic entities by applying relations between signs and meaning. For this purpose, classes of objects, called frames, have been introduced [3]. Frames are structures identified by a name and a set of role-value pairs called slots. Procedures can be attached to slots. A frame can be seen as an organization of concepts. Many deep semantic representations are based on deep case n-ary relations between concepts as proposed by Fillmore and used in the FrameNet project [3]. Deep case systems have very few cases, with each one representing a basic semantic constraint. Frames used in semantics are inspired by case structures [7] and have been considered as cognitive structures. In [14], a theory is presented in which semantic structures are obtained by composition functions. Thematic roles are part of the structures. FrameNet has the goal of documenting the syntactic realization of arguments of the predicates of the general English lexicon by annotating a corpus with semantic roles. The project is focused on task-independent semantic frames, which are defined as a schematic representation of situations involving various participants. First, frames and their participant frame elements are designed, and then example sentences for the frame from the British National Corpus are manually annotated.

Consider the example sentence *The customer accepts the contract*. In order to represent intension, an action expressed by a verb “accept” is represented by a frame as follows:

```
{accept
                is_a      :        verb
                agent              [human.....]
                theme              [.................]
                ........................
                Other roles.......... [.............]}
```

A specific MRL semantic defines this frame as a prototype based on which many instances can be obtained. In these instances, values are associated to roles. Between brackets are represented constraints for the values of each role. Specific representations can be attached to slots such as mass terms, adverbial modification, probabilistic information, degree of certainty, time, and tense. Finding values for roles can be seen as a slot-filling process performed by attached procedures.

PropBank is another project that adds a layer of predicate argument information or semantic role labels to the syntactic structures of the Penn Treebank. Semantic role labels indicate the role of each argument for all target predicates (verbs) in a sentence. For instance, [18] presents a method to bootstrap SLU systems based on PropBank parses. Predicate/argument sets contribute to form a frame when the resulting structure has a specific meaning. For some applications, the only useful composition is a frame representing dialog acts whose components are semantic constituents. Application goals can be represented by frames that constrain the aggregation of predicate/argument pairs to specify system actions.

FROM WORD SEQUENCES TO SEMANTIC INTERPRETATION

Semantic interpretation is based on the application of relations between signs and meaning. The process can be seen as a translation leading to sentences in MRL.

[OBTAINING MEANING FROM SPEECH IS A COMPLEX PROCESS AND MANY DIFFERENT APPROACHES AND MODELS HAVE BEEN PROPOSED.]

An important question concerns the definition of the signs. If it is assumed that signs are words, then the language to be translated is natural language. Some SLU systems are based on this assumption. In this case, interpretation of spoken language is similar to interpretation of written language. A semantic analyzer may interact with a syntactic analyzer to produce semantic representations acceptable to a logical deductive system. This is motivated by arguments that each major syntactic constituent of a sentence maps into a conceptual constituent, but the inverse is not true.

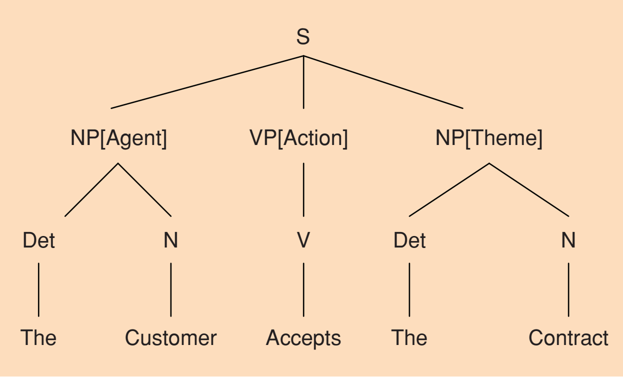
SEMANTIC AND SYNTACTIC ANALYSIS

Syntactic analysis can be performed by a parser that produces a parse tree for a sentence and semantic labels, like predicate and arguments, which can be attached to components of the parse tree. For example, parsing the sentence *The customer accepts the contract* results in the parse tree shown in Figure 1 to which semantic labels are attached.

There are conceptually different approaches for associating semantic labels to nonterminal symbols.

Feature structures may be used to provide constraints on grammatical rules and rule out inadmissible combinations; for example, ensuring agreement between the person and number features of the subject noun phrase (NP) and the verb phrase (VP). In the following sentence, for example, the syntactic features of *who* as opposed to *which* produce a different syntactic analysis:

List all employees of the companies who/which are based in the city center.



[FIG1] **Parse tree with attached semantic labels.**

1) list employees based in the city center

2) list employees of those companies that are based in the city center.

An association of semantic building formulas with syntactic analysis is proposed in categorical grammars (reviewed in [6]) conceived for obtaining a surface semantic representation. The syntax of a language is seen as algebra and grammatical categories are seen as functions. Lexical representations have associated a syntactic pattern that suggests possible continuations of the syntactic analysis and the semantic expression to be generated. Semantic knowledge is associated in this case with lexical entries, and logic formulas are composed by actions performed during parsing. A detailed discussion of rules for deriving semantic interpretations from syntactic parse trees can be found in [1].

DEEP SEMANTIC PROCESSING

A deeper semantic analysis may also require contextual information to recognize and distinguish between different user intentions. In the following example (from [2]), the query:

Can we remove the people by helicopter?

uttered in the context of a disaster management scenario could represent two possible speaker intentions:

1) a request to change the plan (i.e., can we use a helicopter rather than a truck?)

2) a question about feasibility (i.e., is it possible to use a helicopter?).

Further intentions are possible at the problem-solving level; for example,

1) to introduce a new goal

2) to elaborate or extend a solution to the current problem

3) to suggest a modification to the current solution (for example, moving them by truck).

Determining these intentions requires reasoning about the task and current context to identify the most plausible interpretation. The semantic representation used in [2] is a logical form language that provides a domain-independent, unscoped semantic representation of the utterance that can be linked to domain-specific knowledge to support advanced discourse processing that takes context into account.

One problem with richer semantic representations is that, while they have high precision and support deeper understanding, they are usually hand-crafted and suffer from a lack of robustness and efficiency.

Recently a number of resources have become available that support developers of advanced grammatical formalisms. GF Resource Grammar Library [5] enables developers to write grammars using reusable software libraries. The Regulus platform [23] is a development environment for building grammar-based speech applications, supporting the compilation of typed unification grammars into parsers, generators, language models, and recognition packages.

TASK-DEPENDENT SLU MODELS

Many problems of automatic interpretation in SLU systems arise from the fact that many sentences are ungrammatical, the ASR components make errors in hypothesizing words, and grammars have limited coverage. These considerations suggest that it is worth considering specific but more robust models for each conceptual constituent.

In the early 1990s, the Defense Advanced Research Projects Agency (DARPA)-funded Airline Travel Information System (ATIS) project resulted in a number of task-dependent SLU systems. Data were collected with system-user inquiries about flight information; for example, *I want to fly to Boston from New York next week* or *Does this flight serve meals?* ATIS provided a benchmark for many grammar-based, statistical, and hybrid spoken-language understanding systems. The frame elements typically contain information about the departure and arrival cities and date.

[DESIGNING A MEANING REPRESENTATION THAT CAN CAPTURE THE RICH EXPRESSIVITY OF SPOKEN LANGUAGE IS DIFFICULT.]

The linguistic analyzer TINA [26] was proposed by MIT. It is written as a set of probabilistic context-free rewrite rules with constraints, which is converted automatically at run time to a network form in which each node represents a syntactic or semantic category. The probabilities associated with rules are calculated from training data and serve to constrain search during recognition (without them, all possible parses would have to be considered). A robust matcher was obtained by modifying the grammar to allow partial parses. In robust mode, the parser proceeds left to right as usual, but an exhaustive set of possible parses is generated starting at each word of the utterance.

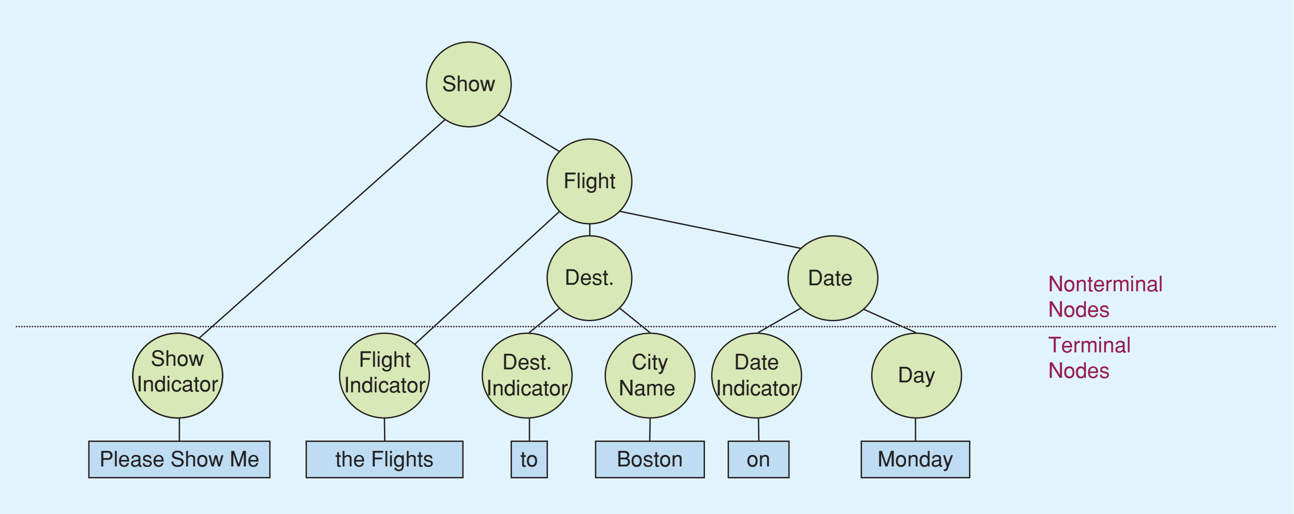
PROBABILISTIC SLU MODELS

In addition to partial parsing and back-off to special matchers when parsing fails (a review can be found in [6] for the ATIS project), it was found useful to construct devices for representing knowledge whose imprecision is characterized by probability distributions. It was also found useful to obtain model parameters by automatic learning using manually annotated corpora. This works as long as manual annotation is easy, reliable, and of sufficiently wide coverage.

The AT&T Chronus SLU system [20] is based on the noisy channel paradigm, commonly used for formalizing the general speech recognition problem. In this system, semantic knowledge is represented by a Markov model in which observations are words w, using one state for each semantic concept.

A tree-structured meaning representation was proposed in the Hidden Understanding Model (HUM) [16]. An example of this representation is shown in Figure 2. Semantic constituents corresponding to partial parse trees are used as names for nonterminal symbols of a stochastic context-free grammar (SCFG). The semantic language model employs tree-structured meaning representations: concepts are represented as nodes in a tree, with subconcepts represented as child nodes. Interpretation is guided by a strategy represented by a stochastic decision tree. Each terminal node is the parent of a word or of a sequence of words. It is worth noticing that full parsing of a sentence requires the sentence bounds to be known.

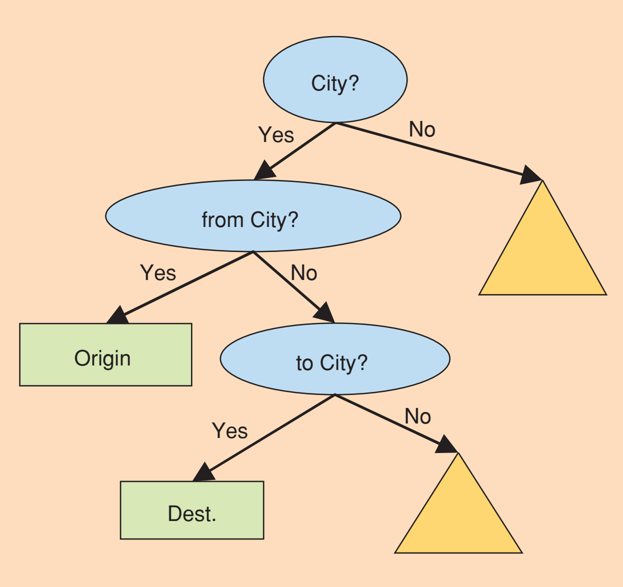
AT&T Chronus and BBN HUM systems are reviewed in detail in [28] where a hybrid approach with Markov models and SCFG is also described. In this approach, a manually built CFG is exploited for embedding domain-specific and domain-independent knowledge, like a city name list and a date grammar. The CFG rules can be populated with database entries or prebuilt in a grammar library for domain-independent concepts (e.g., date and time).



[FIG2] An example hierarchical semantic representation for the ATIS domain (from [16]).

In the Chanel system, first the general entities, such as city or airport names and dates, are marked. Then their roles are determined using semantic classification trees (SCTs) automatically learned using the training data. A simplified example decision tree is shown in Figure 3 for assigning the role of a city. Frame instances are composed by rules from role hypotheses. More details can be found in [6, chap. 14].

At Cambridge University [13], an approach based on SCFGs was proposed that does not require fully annotated data for training. The proposed solution considers a hidden vector state (HVS) model. Each state encodes the tree context in a vector (or a stack). Then the state transition for each emitted word is factored into n stack pop operations and one push operation for the word itself. Therefore, the model can only represent right-branching context-free grammars. The transition probabilities for each n are learned from the corpus or from the grammar templates. The stack depth and number of pop operations n are restricted for efficiency reasons.

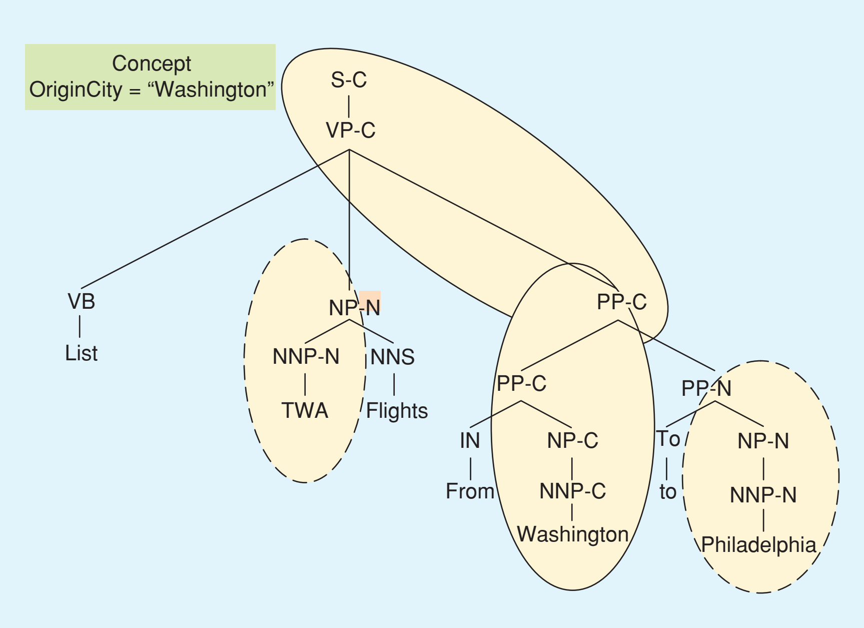


[FIG3] Example of an SCT for the ATIS domain.

The use of stochastic semantic grammars at Microsoft is reviewed in [28].

More recently, combined statistical models of syntax and semantics have been proposed. Figure 4 shows a syntactic-semantic tree for the sentence *List the TWA flights from Washington to Philadelphia*. While the parse has been extracted from a generic English parser, its subtrees have been marked semi-automatically with features relevant to the attribute OriginCity (full line) and contextual (dashed line) to the sentence [17] . With this approach, semantic hypotheses are dynamically attached to nonterminal symbols of a general-purpose syntactic grammar, rather than having a semantic grammar with static rules defining possible rewriting of semantic nonterminal symbols.

In [19] statistical translation models are used to translate a source sentence S into a target MRL T by maximizing the probability $P(T|S)$. The central task in training is to determine correlations between groups of words in one language and MRL entities. The source channel fails to capture such correlations, so a model has been built to directly compute the posterior probability $P(T|S)$. The use of discriminative models for this purpose is described in [30].



[FIG4] Example of a syntactic-semantic tree for the ATIS domain.

CLASSIFICATION MODELS FOR SLU

Robust spoken language understanding has been addressed in the context of dialog utterance or user intent classification. Pioneering work was done in AT&T's How May I Help You? [8], where the users' conversational spoken utterances are classified into a number of predefined intents. The approach taken uses a discriminative classifier (boosting or SVM) with the word n-grams and task-specific named entities appearing in the utterance as features [9]. Combinations of human-crafted knowledge and the results of automatic learning of semantic knowledge are proposed in [25]. The concurrent use of SCT, boosting, and SVM classifiers is proposed in [22] to increase classification robustness.

CORPUS COLLECTION AND COMPARISON OF DIFFERENT APPROACHES

Regardless of the method chosen for performing the translation process from a speech signal to a semantic interpretation, corpora are needed to perform system development and evaluation. For manually defined rule-based systems, corpora are useful for writing the rules through the study of speech samples and by checking their coverage. For corpus-based methods, corpora with semantic annotations are used to train models. For evaluation purposes both approaches need corpora with manual semantic annotations as references.

The availability of speech corpora for SLU is a major issue due to the difficulties involved in building a semantic model as well as the costs of collecting and manually annotating a speech corpus with the semantic information required for such a model.

[THE SYNTAX OF A LANGUAGE IS SEEN AS ALGEBRA AND GRAMMATICAL CATEGORIES ARE SEEN AS FUNCTIONS.]

The corpora linked to a specific application are generally not publicly available. However, two notable examples of publicly available semantically annotated dialog speech corpora are the ATIS corpus mentioned previously and the French MEDIA corpus, a spoken dialog corpus about tourist enquiries collected using the Wizard of Oz method. Description, discussion, experiments, and results can be found in [6, chapter 14] for the first corpus and in [30] for the second corpus. There is not a single approach that is a clear winner from the SLU evaluations performed on these corpora. Statistical methods have proven to be more robust at the cost of having a large annotated training corpus, which is not always feasible. Among the statistical methods proposed for SLU, discriminative methods such as conditional random fields provide an advantage over generative methods, as reported in [30] on the MEDIA corpus. However, as pointed out in the same paper, discriminative methods seem to be more affected by errors and lack of consistency in the training corpus annotations than generative methods.

DEALING WITH MULTIPLE ASR HYPOTHESES

One way of dealing with speech recognition errors is to take into account not only the best word string obtained thanks to ASR models but a set of multiple hypotheses that can be represented as a word lattice or an n-best list. When lattices of word hypotheses are generated, it is likely that the uttered words are hypothesized somewhere in the lattice, making it possible to obtain coherent semantic hypotheses from selected signs. Algorithms have been proposed for generating probabilistic lattices of conceptual constituent hypotheses from a probabilistic lattice of word hypotheses or during speech decoding.

In [22] the conceptual decoding process is seen as a translation process in which stochastic language models are implemented by finite state machines (FSMs) that output labels for semantic constituents. There is an FSM for each elementary conceptual constituent. Each FSM implements a finite state approximation of a natural language grammar. These FSMs are transducers that take words as input and output the concept tag conveyed by the accepted phrase. At decoding time they are applied to the word graphs output by the ASR decoder by means of a composition operation. This word/concept lattice is then rescored using an HMM-based concept tagger. This approach is called an integrated decoding approach as the ASR and SLU processes are done together by looking at the same time for the best sequence of words and concepts.

Another example of an integrated approach that takes as input a word lattice is proposed at AT&T in [4]. A mixture language model for a multimodal application is described with a component trained with in-domain data and another obtained with data generated by a grammar. Understanding is the recognition of the sequence of predicate/argument tags that maximize $P(T|W)$ where T is the tag sequence and W is the sentence. An approximation is made by considering bigrams and trigrams of tags.

At IBM [24], a system is proposed that generates an n-best list of word hypotheses with a dialog state-dependent trigram **language model (LM)** and rescores them with two semantic models. An embedded context-free semantic grammar (EG) is defined for each concept that performs concept spotting by searching for phrase patterns corresponding to concepts. Trigram probabilities are used for scoring hypotheses with the EG model. A second LM, called maximum entropy LM (MELM), computes probabilities of a word, given the history, using an ME model.

SEMANTIC CONFIDENCE

Current state-of-the-art ASR and SLU systems make errors that have to be identified in order to apply appropriate strategies for performing communicative and system actions, such as error correction and repair in human-computer dialogs. The posterior probability $P(I|Y)$ of an interpretation I given a time sequence of acoustic features Y is not the best reliability indicator for a hypothesis as suitable confidence indices should also take into account information that is not coded in Y, such as the coherence of the available hypotheses with the entire dialog history, including system prompts and repairs.

Estimating the confidence of an interpretation raises several issues: choosing the span of the confidence measures (word, conceptual constituent, or utterance), defining the set of features involved in the confidence estimation (ASR features, SLU features, dialog context), combining efficiently the different features, and choosing a decision strategy that takes into account all the features obtained.

CONFIDENCE AT THE WORD, CONCEPT, AND UTTERANCE LEVELS

Two levels of features to train confidence models for words are proposed in [12]. They are word-level features that focus only on the reliability of acoustic samples, and utterance-level features that concern the appropriateness of the whole utterance in which the word is found. The assumption is made that if the whole utterance is unreliable, then the word contained in that utterance is likely to be incorrect.

Confidence measures based on N-best lists and content words are defined as the sum of the posterior probabilities of sentences in an N-best list containing the content word. Posterior probabilities can be obtained from word graphs and concept graphs as in [10].

COMBINING CONFIDENCE FEATURES

With the purpose of combining multiple knowledge sources at different levels, in [24] previous approaches to the integration of semantic and other ASR features are reviewed and it is observed that, in most cases, their integration into the decision process is rather ad hoc. Word and concept-level confidence annotations are considered. Two methods are proposed that use two sets of statistical features to model the presence of semantic information in a sentence. The first relies on a semantic tree where node and extension scores are used. Scores are based on the assumption that sentences that are grammatically correct and likely to be free of recognition errors tend to be easier to parse and should receive high confidence. The second technique is based on joint ME modeling of the words of a sentence and the semantic parse tree.

ONE PROBLEM WITH RICHER SEMANTIC REPRESENTATIONS IS THAT THEY ARE USUALLY HAND-CRAFTED AND SUFFER FROM A LACK OF ROBUSTNESS AND EFFICIENCY.

INTEGRATING DIALOG CONTEXT

In spoken dialog systems, it is important to use confidence measures that integrate information related to the whole dialog context rather than just having features based only on acoustic and language model cues. The integration of dialog manager expectations is proposed in [21]. The dialog expectations are represented by clusters of dialog prompts that are used as features, in conjunction with acoustic and linguistic features, in a decision tree trained to assign confidence to a semantic interpretation.

SLU AND DIALOG STRATEGIES BASED ON CONFIDENCE SCORES

In [22] an interpretation strategy implemented by a decision tree is proposed. In addition to confidence measures given at the concept level, a limited set of reliability states is defined for characterizing a whole utterance interpretation. The set of features used involves semantic information through semantic classifiers, linguistic and acoustic information, and dialog expectations. Some semantic confidence indicators are based on the agreement of semantic interpretations obtained by different classification methods. The reliability state of a hypothesis corresponds to a global confidence measure. The kinds of interpretation errors that are expected in each state can also be predicted and an error correction strategy that can reject or add conceptual constituents to the best interpretation obtained in the first stage of the SLU process is proposed. Dialog managers with error-handling strategies based on confidence measures have also been proposed. In [27], a Markov decision processes (MDP) approach is used where concepts are represented by partially observable MDPs, with three underlying hidden states: correct, incorrect, and empty. The belief state is constructed at each time step from the confidence score of the top hypothesis for the concept.

ADAPTIVE LEARNING

The SLU prediction models described in the previous sections require corpora with annotations $(\mathbf{X}, \mathbf{Y})$ where the observation vector $\mathbf{X}$ is aligned with a label vector $\mathbf{Y}$. For example, $\mathbf{X}$ corresponds to the sequence of words output by the ASR and $\mathbf{Y}$ is expected to be the aligned sequence of semantic units or concepts. The approaches described so far for SLU use algorithms that process the observation $\mathbf{X}$ (e.g., words) and $\mathbf{Y}$ in batch mode (i.e., infinite time horizon). In the context of real-time spoken dialog systems, online learning models of SLU are required. In this class of learning algorithms the statistical parameters of the model are updated on a sample-by-sample basis and are actively learned. A constraint is applied on the time horizon of the learning model. SLU architectures with incremental learning are crucial to mimic human performance in sentence processing. This is certainly an area of research in its infancy and will likely receive more attention in next-generation spoken dialog systems. Active learning is a component of the adaptive prediction model that is able to select samples that are more likely to improve its performance. There have been important results in so-called batch-mode active learning motivated by the cost of manually annotating $\mathbf{X}$ with $\mathbf{Y}$ for training SLU models. Some open issues in the development of such models are how to collect a semantically annotated corpus while the system is not deployed and how off-the-shelf corpora from different applications can be used in the training process. Even when unlabeled data are available, manual annotation of a large number of utterances is labor intensive and time consuming. To alleviate this problem, active and unsupervised learning mechanisms have been proposed [11]. Active learning aims to minimize the number of labeled utterances by automatically selecting the utterances that are likely to be the most informative for annotation. The unlabeled examples can further be exploited using semi-supervised learning methods. In the machine learning literature certainty-based and committee-based selective sampling methods have been proposed for active learning [11]. It was observed that there is a reverse correlation between the confidence score given by the classifier used in these methods and the informativeness of that utterance to be interpreted. That is, the higher the SLU confidence scores for an utterance, the less informative is that utterance. Active learning for intent determination resulted in a 75% reduction of the amount of manually labeled data needed for a large-scale call classification application. Note that the semantic confidence score estimation presented in the previous section has direct implication on active and semi- or unsupervised learning.

The process of learning is coupled with the annotation lexicon ($\mathbf{Y}$) and the evaluation metrics. The output label set $\mathbf{Y}$ should be such that it is nonambiguous (with respect to human annotation error), easy to acquire (with respect to cost and time), and effective in terms of learning rate. In the ATIS project the so-called common answer specification (CAS) metric was designed to associate a natural language query to a set of answers generated by a predefined syntax. Such an annotation model was expensive both in terms of time and cost, as well as oriented to a database query task. More recently, a simpler and domain-dependent model of concept stream associated to each spoken query has been adopted. In this case, $\mathbf{Y} = (y_1, \ldots y_M)$, where y_i belongs to the set of predefined labels that could be domain dependent or independent. The evaluation metric in this case is the concept error rate (CER), which is a valuable tool for system development and evaluation. In the latter case the burden is on the annotation designer to select a set of concepts (and its ontology), which is stable (with respect to annotator agreement), scalable, and portable to other domains. SLU evaluation is still an open research issue and is tightly coupled with the evaluation of conversational systems. The ability to evaluate SLU in the context of spoken dialog systems will be one of the key aspects of next-generation SLU interfaces.

Another challenging aspect of SLU for third-generation conversational systems is the ability to learn the semantic interpretation drawing from a priori knowledge models as well as grounding it in the physical/virtual world. Machines should be able to leverage from the speech channel as well as from multimodal signs (e.g., pen gestures, visual cues, etc.), following an incremental understanding process.

CONCLUSIONS AND FUTURE PERSPECTIVES

SLU is one of the fundamental processes in spoken communication. Among the many important related research problems, it is worth mentioning the study of computational models of semantic theories, the decision process about interpretation with imprecise sources of knowledge, and imprecise decoding of signs from the speech signal. Decisions about interpretation have to be made by minimizing the risk of errors. As models and hypotheses are imprecise, probability distributions have to be associated with them. Evaluation of the confidence of interpretation hypotheses is important for the strategy that has to use them. Interpretation knowledge is based on knowledge that can be partially or totally acquired from annotated corpora. Semantic annotation is costly, and methods for incremental learning are of fundamental importance. Finally, research on SLU has the potential to create technology for (partial) automation of call centers providing services such as customer care, help desk, and opinion analysis.

Major directions for future research in SLU are: meaning representation, the definition and representation of signs, the conception of relations between signs and meaning and between instances of meaning, processes for sign extraction, generation of hypotheses about units of meaning (also called semantic constituents), and constituent composition into semantic structures. As processes generate interpretation hypotheses, other challenging problems are the robustness and evaluation of confidence for semantic hypotheses, the design of interpretation knowledge sources using human knowledge, automatic learning of relations from annotated corpora, and the collection and semantic annotation of corpora with limited human effort.

ACKNOWLEDGMENTS

This work was partially supported by the European Union, Project LUNA, IST contract no. 33549 and the Marie Curie Excellence Grant for the ADAMACH project (contract 022593).

AUTHORS

Renato De Mori (renato.demori@univ-avignon.fr) received a doctorate degree in electronic engineering from Politecnico di Torino (Italy). He is a Fellow of the IEEE Computer Society and distinguished lecturer of the IEEE Signal Processing Society. He is the author or editor of four books and has published more than 100 scientific papers in many international journals. His major contributions have been in the areas of automatic speech recognition, computer arithmetic, software engineering, and human-machine interfaces. He was chief editor of *Speech Communication* and associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He is on the editorial board of *Computer Speech and Language*. He has been on the editorial board of *Computational Intelligence*, *Pattern Recognition Letters*, and *Signal Processing*. He was a member of the Executive Advisory Board at the IBM Toronto Lab, scientific advisor at France Telecom R&D, chairman of the Computer and Information Systems Committee of the Natural Sciences and Engineering Council of Canada, and vice-president of R&D for the Centre de Recherche en Informatique de Montreal. He has been a member of the IEEE Speech Technical Committee, the Interdisciplinary Board of the Canadian Foundation for Innovation, and the Interdisciplinary Committee for Canadian chairs.

Frederic Béchet (frederic.bechet@lia.univ-avignon.fr) received the Ph.D. in computer science from the University of Avignon, France, in 1994. Since 1995 he has been an assistant professor and a researcher at the computer laboratory of the University of Avignon (LIA). He is the author/coauthor of over 50 refereed papers in journals and international conferences. He was an invited professor at the AT&T Research Lab in Florham Park, New Jersey, from August 2001 until September 2002. His research activities focus mainly on spoken language understanding in human-computer dialog context. He is a member of ISCA, ACL, IEEE Signal Processing Society, and the Speech and Language Technical Committee (SLTC) for 2008–2010.

Dilek Hakkani-Tür (dilek@icsi.berkeley.edu) is a senior researcher at ICSI. Prior to that, she worked for AT&T Labs—Research. She received her B.Sc. degree from Middle East Technical University in 1994 and M.Sc. and Ph.D. degrees from

Bilkent University, Computer Engineering Department, in 1996 and 2000, respectively. Her research interests include spoken language processing, spoken dialog systems, and active and unsupervised learning. She is a member of ISCA, IEEE, ACL, and an associate editor of the *IEEE Transactions on Audio, Speech and Language Processing*.

Michael McTear (mf.mctear@ulster.ac.uk) is a professor of knowledge engineering at the University of Ulster, Northern Ireland, where he has been an active researcher in spoken language technologies for more than 15 years. He is author of the book *Spoken Dialogue Technology: Toward the Conversational User Interface* (Springer Verlag, 2004) and has been involved in a number of research projects in spoken dialog technology. In December 2006 he presented an invited tutorial at the IEEE/ACL Workshop on Spoken Language Technology in Aruba and in June 2007 he presented an invited tutorial on Spoken Dialogue Systems at the ACL conference in Prague.

Giuseppe Riccardi (riccardi@dit.unitn.it) is a professor in the faculty of Electrical Engineering at University of Trento (Italy). He is member of the EECS Department and the Center for Mind/Brain Sciences. He is the director of the Adaptive Multimodal Information and Interfaces Lab. In 2006 he was awarded by the European Commission the Marie Curie Research Excellence Leader position. From 1993–2005, he was with AT&T Bell Laboratories and then AT&T Labs-Research. He has coauthored more than 80 papers and 25 patents in the field of automatic speech recognition, spoken language understanding, machine translation and spoken dialog systems. He has been on the scientific committee of many international conferences and received many international awards. He is on the editorial board of the *ACM Transactions of Speech and Language*. He is member of the IEEE SLTC Speech Technical Committee. He is a Senior Member of the IEEE, ACL and ISCA.

Gokhan Tur (gokham@speech.sri.com) received his B.S., M.S., and Ph.D. degrees from the Department of Computer Science, Bilkent University, Turkey in 1994, 1996, and 2000, respectively. He worked at AT&T Labs—Research from 2001 to 2006. He is currently with the Speech Technology and Research Lab of SRI International. His research interests include spoken language understanding, speech and language processing, machine learning, and information retrieval and extraction. He coauthored more than 60 papers published in refereed journals and presented at international conferences. He received the *Speech Communication Journal* Best Paper awards by ISCA for 2004–2006 and by EURASIP for 2005–2006. He was the organizer of the HLT-NAACL 2007 Workshop on Spoken Dialog Technologies and the HLT-NAACL 2004 and AAAI 2005 Workshops on Spoken Language Understanding. He was the spoken language processing area chair for IEEE ICASSP 2007 and IEEE ICASSP 2008 conferences, spoken dialog area chair for the HLT-NAACL 2007 conference, finance chair for the IEEE/ACL SLT 2006 Workshop, and spoken language understanding area chair for the IEEE ASRU 2005 workshop. He is a member of IEEE Signal Processing Society Speech and Language Technical Committee (SLTC) for 2006–2008.

REFERENCES

[1] J. Allen, *Natural language Understanding*. Menlo Park CA: Benjamin/Cummings, 1987.

[2] J. Allen, M. Dzikovska, M. Manshadi, and M. Swift, "Deep linguistic processing for spoken dialogue systems," in *Proc. ACL Workshop on Deep Linguistic Processing*, Prague, Jun. 2007, pp. 49–56.

[3] C. Baker, C.J. Fillmore, and J. Lowe, "The Berkeley FrameNet project," in *Proc. COLING-ACL*-1998, pp. 86–99.

[4] S. Bangalore and M. Johnston, "Balancing data-driven and rule-based approaches in the context of a multimodal conversational system," in *Proc. IEEE Workshop on Automated Speech Recognition and Understanding*, Nov. 2003, pp. 221–226.

[5] B. Bringert, "Speech recognition grammar compilation in Grammatical Framework," in *Proc. ACL Workshop on Grammar-Based Approaches to Spoken Language Processing*, Prague, Jun. 2007, pp. 1–8.

[6] R. De Mori, *Spoken Dialogues with Computers*. New York: Academic, 1998.

[7] C.J. Fillmore, "The case for case" in *Universals in Linguistic Theory*, E. Bach and R. Harms, Eds. New York: Holt, Rinehart and Winston, 1968.

[8] A.L. Gorin, G. Riccardi, and J.H. Wright, "How may I help you?" *Speech Commun.*, vol. 23, pp. 113–127, Oct. 1997.

[9] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Gilbert, "The AT&T Spoken Language Understanding System," *IEEE Trans. Speech Audio Processing* vol. 14, no. 1, pp. 213–222. 2006.

[10] K. Hacioglu and W. Ward, "A concept graph based confidence measure," in *Proc. ICASSP 2002*, Orlando, FL, vol. 1, pp. 225–228.

[11] D Hakkani-Tür, G. Riccardi, and G. Tur, "An active approach to spoken language processing," *ACM Trans. Speech Language Processing*, vol. 3, no. 3, pp. 1–31, 2006.

[12] T.J. Hazen, S. Seneff, and J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems," *Comput. Speech Lang.*, vol. 16, pp. 49–68, Jan. 2002.

[13] Y. He and S. Young, "Spoken language understanding using the Hidden Vector State Model," *Speech Commun.*, vol. 48, pp. 262–275, Feb. 2006.

[14] R. Jackendoff, *Foundations of Language*. Oxford, UK: Oxford Univ. Press, 2002.

[15] M. McTear, "Spoken language understanding for conversational dialog systems," in *Proc. IEEE/ACL Workshop on Spoken Language Technology.* Palm Beach, Aruba, Dec. 2006, p. 6.

[16] S. Miller, R. Schwartz, R. Bobrow, and R. Ingria, "Statistical language processing using hidden understanding models," in *Proc. Workshop on Human Language Technology*, Plainsboro, NJ, 1994, pp. 278–282.

[17] A. Moschitti, G. Riccardi, and C. Raymond, "Spoken language understanding with kernels for syntactic/semantic structure," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Kyoto, Dec. 2007, pp. 183–188.

[18] G. Tur, D. Hakkani-Tür, and A. Chotimongkol, "Semi-supervised learning for spoken language understanding using semantic role labeling," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Puerto Rico, Dec. 2005, pp. 232–237.

[19] K.A. Papieni, S. Roukos, and R.T. Ward, "Maximum likelihood and discriminative training of direct translation models," in *Proc. IEEE ICASSP*, Seattle WA, May 1998, pp. 189–192.

[20] R. Pieraccini, E. Levin, and C.H. Lee, "Stochastic representation of conceptual structure in the ATIS task," in *Proc. Workshop on Speech and Natural Language*, Los Altos, CA, 1991, pp. 121–124.

[21] A. Potamianos, S. Narayanan, and G. Riccardi, "Adaptive categorical understanding for spoken dialogue systems," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 2, pp. 321–329, 2005.

[22] C. Raymond, F. Béchet, N. Camelin, R. De Mori, and G. Damnati, "Sequential decision strategies for machine interpretation of speech," *IEEE Trans. Speech Audio Processing*, vol. 15, vol. 1, pp. 162–171, 2007.

[23] M. Rayner, B.A. Hockey, and P. Bouillon, *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. Chicago: Univ. of Chicago Press, 2006.

[24] R. Sarikaya, Y. Gao, M. Picheny, and H. Erdogan, "Semantic confidence measurement for spoken dialog systems," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 4, pp. 534–545. 2005.

[25] R.E. Schapire, M. Rochery, M. Rahim, and N. Gupta, "Boosting with prior knowledge for call classification," *IEEE Trans. Speech Audio Processing,* vol. 13, no. 2, pp. 174–182, 2005.

[26] S. Seneff, "TINA: A probabilistic syntactic parser for speech understanding systems," in *Proc. IEEE ICASSP*, Glasgow, UK, 1989, vol. 2, pp. 711–714.

[27] D. Bohus and A. Rudnicky, "Error handling in the RavenClaw dialog management architecture," in *Proc. HLT-EMNLP 2005*, Vancouver, CA, pp. 225–232.

[28] Y.-Y. Wang; L. Deng; and A. Acero, "Spoken language understanding," *IEEE Signal Processing Mag.*, vol. 22, no. 5, pp. 16–31, 2005.

[29] W.A. Woods "What's in a link?" in *Representation and Understanding*, D.G. Bobrow and A. Collins, Eds. New York: Academic, 1975.

[30] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Proc. Interspeech,* Antwerp, Belgium, Aug. 2007, pp. 1605–1608.