1

Introduction

Gokhan Tur[1] and Renato De Mori[2]
[1] *Microsoft Speech Labs, Microsoft Research, USA*
[2]*McGill University, Canada and University of Avignon, France*

COPYRIGHTED MATERIAL

1.1 A Brief History of Spoken Language Understanding

In 1950, Turing published his most cited paper, entitled "Computing Machinery and Intelligence", trying to answer the question "Can machines think?" (Turing, 1950). Then he proposed the famous imitation game, or the Turing test, which tests whether or not a computer can successfully imitate a human in a conversation. He also prophesied that "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted". Yet, now we are well past the year 2000, and we wonder whether he meant the end of 21st century when machines will be able to "understand" us.

Spoken language understanding (SLU) is currently an emerging field in the intersection of speech processing, natural language processing (NLP) by leveraging technologies from machine learning (ML) and artificial intelligence (AI). While speech is the most natural medium people use to interact with each other, when using tools, machines, or computers we use many other modalities such as mouse, keyboard, stylus, but not speech. Similarly, when people talk to each other, there is no record and the words are simply lost. However there is strong interest – both commercial and academic – in understanding such communications. As speech recognition and NLP algorithms mature, these areas become far from unreachable dreams. It is clear that we will see an increase in the number of SLU applications in the future. For example, robots will better understand what we say instead of reacting to a finite number of predefined commands, or we will be able to ask the computer to retrieve a phone call with a certain person about a particular topic.

Simply put, language understanding aims to extract "meaning" from natural language. In the literature on social sciences and artificial intelligence, many great names have worked on formalizing what meaning constitutes. Jackendoff, who has worked for many decades on the

Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, First Edition.
Edited by Gokhan Tur and Renato De Mori.
© 2011 John Wiley & Sons, Ltd. Published 2011 by John Wiley & Sons, Ltd.

semantics of natural language and its lexical and syntactic expressions, admits that "meaning" is the "holy grail" for linguistics and philosophy (Jackendoff, 2002).

One favorite quote from Shannon's well-known 1948 paper on information theory says that: *The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning. . . . These semantic aspects of communication are irrelevant to the engineering problem* (Shannon, 1948). It is ironic that information-theory-based approaches now dominate the field of spoken language understanding. Apparently, and interestingly using such data-driven approaches is very well suited for processing spoken language after several decades of over-promised and under-delivered attempts by using only knowledge-based approaches which tried to mimic the human brain for building intelligent agents. As one of the pioneers of statistical spoken language processing, Fred Jelinek, once put it: *The human analogy is misguided because machines do things differently than biological beings. Airplanes don't flap their wings*.

People no longer believe that machines will be able to converse like humans in the near future. On the other hand, a number of systems aimed at understanding have been developed in the last decade. It is possible to categorize these systems into three, based on their high-level approaches. The first category is the ELIZA like systems, where the system does not really understand what is said, but just mimics understanding. ELIZA was built at MIT in the 1960s (Weizenbaum, 1966) and has been the most popular artificial intelligence program, even though there was no intelligence but instead relied on simple pattern matching. A typical interaction may have such excerpts:

User: I need some help

System: Why do you say you need some help?

User: I am depressed

System: Are you depressed often?

Contemporary clones of ELIZA, such as ALICE,[1] are moving towards embedding more sophisticated language processing technologies within the same framework.

The understanding systems in the second category are rooted in artificial intelligence. They are demonstrated to be successful for very limited domains, using deeper semantics. These systems are typically heavily knowledge-based and rely on formal semantic interpretation, defined as mapping sentences into their logical forms. In its simplest form, a logical form is a context-independent representation of a sentence covering its predicates and arguments. For example, if the sentence is "John loves Mary", the logical form would be "(love john mary)".

During the 1970s, the first systems for understanding continuous speech were developed with interesting approaches for mapping language features into semantic representations. For this purpose, *case grammars* were proposed for representing sets of semantic concepts with thematic roles such as agent or instrument. The ICSI FrameNet project, for instance, focused on defining semantic frames for each of the concepts (Lowe and Baker, 1997). For example, in the "commerce" concept, there is a "buyer" and a "seller" and other arguments such as the "cost", "good", and so on. Therefore, two sentences "A sold X to B" and "B bought X from A" are semantically parsed as the same. Following these ideas, some researchers worked towards building universal semantic grammars (or interlingua), which assumes that all languages have a shared set of semantic features (Chomsky, 1965). Such interlingua-based approaches have

[1] http://alicebot.blogspot.com/

also heavily influenced language translation until late 1990s before statistical approaches began to dominate the field. Allen (1995) may be consulted for more information on the artificial-intelligence-based techniques for language understanding.

The last category of understanding systems is the main focus of this book, where understanding is reduced to a (mostly statistical) language processing problem. This corresponds to attacking targeted speech understanding tasks instead of trying to solve the global machine understanding problem. A good example of targeted understanding is detecting the arguments of an intent given a domain, as in the Air Travel Information System (ATIS) (Price, 1990). ATIS was a popular DARPA-sponsored project, focusing on building an understanding system for the airline domain. In this task, the users utter queries on flight information such as *I want to fly to Boston from New York next week*. In this case, understanding was reduced to the problem of extracting task specific arguments in a given frame-based semantic representation involving, for example, *Destination* and *Departure Date*. While the concept of using semantic frames is motivated by the case frames of the artificial intelligence area, the slots are very specific to the target domain, and finding values of properties from automatically recognized spoken utterances may suffer from automatic speech recognition errors and poor modeling of natural language variability in expressing the same concept. For these reasons, the spoken language understanding researchers employed known classification methods for filling frame slots of the application domain using the provided training data set and performed comparative experiments. These approaches used generative models such as hidden Markov models (Pieraccini *et al*., 1992), discriminative classification methods (Kuhn and Mori, 1995) and probabilistic context free grammars (Seneff, 1992; Ward and Issar, 1994).

While ATIS project coined the term "spoken language understanding" for human/machine conversations, it is not hard to think of other interactive understanding tasks, such as spoken question answering, voice search, or other similar human/human conversation understanding tasks such as named entity extraction or topic classification. Hence, in this book, we take a liberal view of the term spoken language understanding and attempt to cover such popular tasks which can be considered under this umbrella term. Each of these tasks are studied extensively, and the progress is fascinating.

SLU tasks aim at processing either human/human or human/machine communications. Typically the tasks and the approaches are quite different for each case. Regarding human/machine interactive systems, we start from the heavily studied tasks of determination of intent of its arguments and their interaction with the dialog manager within a spoken dialog system. Recently question answering from speech has become a popular task for human/machine interactive systems. Especially with the proliferation of smart phones, voice search is now an emerging field with ties to both NLP and information retrieval. With respect to human/human communication processing, telephone conversations or multi-party meetings are studied in depth. Recently, the established language processing tasks, such as speech summarization and discourse topic segmentation, have been developed to process human/human spoken conversations. The extraction of specific information from speech conversations to be used for mining speech data and speech analytics is also considered in order to ensure quality of a service or monitor important events in application domains.

With advances in machine learning, speech recognition, and natural language processing, SLU, in the middle of all these fields, has improved dramatically during the last two decades. As the amount of available data (annotated or raw) has grown with the explosion of web sources and other information kinds, another exciting area of research area is coping with

spoken information overload. Since SLU is not a single technology, unlike speech recognition, it is hard to present a single application. As mentioned before, any speech processing task eventually requires some sort of spoken language processing. Conventional approaches of plugging in the output of a speech recognizer to the natural language processing engine is not a solution in most cases. The SLU application must be robust to speech, speech recognition errors, certain characteristics of uttered sentences, and so on. For example, most utterances are not grammatical and have disfluencies, and hence off-the-shelf syntactic parsers trained with written text sources, such as newspaper articles, fail frequently.

There is also a strong interest from the commercial world about SLU applications. These typically employ knowledge-based approaches, such as building hand-crafted grammars or using a finite set of commands, and are now used in some environments such as cars, call-centers, and robots. This book also aims to fill this chasm in approaches employed between commercial and academic communities.

The focus of the book will be to cover the state-of-the-art approaches (mostly data-driven) for each of the SLU tasks, with chapters written by well-known researchers in the respecive fields. The book attempts to introduce the reader to the most popular tasks in SLU.

This book is proposed for graduate courses in electronics engineering and/or computer science. However it can also be useful to social science graduates with field expertise such as psycholinguists, linguists, and to other technologists. Experts in text processing will notice how certain language processing tasks (such as summarization or named entity extraction) are handled with speech input. The members of the speech processing community will find surveys of tasks beyond speech and speaker recognition with a comprehensive description of spoken language understanding methods.

1.2 Organization of the Book

This book covers the state-of-the-art approaches to key SLU tasks as listed below. These tasks can be grouped into two categories based on their main intended application area, processing human/human or human/machine conversations, though in some cases this distinction is unclear.

For each of these SLU tasks we provide a motivation for the task, a comprehensive literature survey, the main approaches and the state of the art techniques, and some indicative performance figures in established data sets for that task. For example, when template filling is discussed, ATIS data is used since it is already available for the community.

1.2.1 Part I. Spoken Language Understanding for Human/Machine Interactions

This part of the book covers the established tasks of SLU, namely slot filling and intent determination as used in dialog systems, as well as newer understanding tasks which focus on human/machine interactions such as voice search and spoken question answering. Two final chapters, one on describing SLU in the framework of modern dialog systems, and another discussing active learning methods for SLU conclude Part I.

Chapter 2 History of Knowledge and Processes for Spoken Language Understanding

This chapter reviews the evolution of methods for spoken language understanding systems. Automatic systems for spoken language understanding using these methods are then reviewed, building the stage for the rest of Part I.

Chapter 3 Semantic Frame Based Spoken Language Understanding

This chapter provides a comprehensive coverage of semantic frame-based spoken language understanding approaches as used in human/computer interaction. Being the most extensively studied SLU task, we try to distill the established approaches and recent literature to provide the reader with a comparative and comprehensive view of the state of the art in this area.

Chapter 4 Intent Determination and Spoken Utterance Classification

This chapter focuses on the complementary task of semantic template filling tasks, i.e. spoken utterance classification techniques and illustrates their successful applications to intent determination systems which has emerged partly from commercial call-routing applications. We aim to provide details of such systems, the underlying approaches, and integration with speech recognition and template filling.

Chapter 5 Voice Search

This chapter focuses on one of the most actively investigated speech understanding technologies in recent years: querying a database, such as using speech for directory assistance. A variety of applications (including multi-modal) will be reviewed and the proposed algorithms are discussed in detail along with proposed evaluation metrics.

Chapter 6 Spoken Question Answering

This chapter covers question answering from spoken documents, but also beyond this where questions are spoken. Various approaches and systems for question answering, are presented in detail, with a focus on approaches used for spoken language and on the QAst campaigns.

Chapter 7 SLU in Commercial and Research Spoken Dialog Systems

This chapter shows how different SLU techniques are integrated into commercial and research dialog systems. The focus is providing a comparative view based on example projects, architectures, and corpora associated with the application of SLU to spoken dialog systems.

Chapter 8 Active Learning

This chapter reviews active learning methods that deal with the scarcity of labeled data, focusing on spoken language understanding applications. This is a critical area as statistical, data-driven approaches to SLU have become dominant in recent years. We present applications of active learning for various tasks that are described in this book.

1.2.2 Part II. Spoken Language Understanding for Human/Human Conversations

This part of the book covers SLU tasks, which mainly focus on processing human/human spoken conversations such as multi-party meetings, broadcast conversations, and so on. The first chapter serves as a preamble to Part II, since the chapter discusses lower-level tasks, and higher-level SLU applications, such as topic segmentation or summarization are discussed in the following chapters.

Chapter 9 Human/Human Conversation Understanding

This chapter introduces human/human conversation understanding approaches, mainly focusing on discourse modeling, speech act modeling, and argument diagramming. This chapter also serves as a bridge to other higher-level tasks and studies towards processing human/human conversations, such as summarization or topic segmentation.

Chapter 10 Named Entity Recognition

This chapter discusses the major issues concerning the task of named entity extraction in spoken documents. After defining the task and its application frameworks in the context of speech processing, a comparison of different entity extraction approaches is presented in detail.

Chapter 11 Topic Segmentation

This chapter discusses the task of automatically dividing single long recordings or transcripts into shorter, topically coherent segments. Both supervised and unsupervised machine learning approaches, rooted in speech processing, information retrieval, and natural language processing are discussed.

Chapter 12 Topic Identification

This chapter builds on the previous chapter and focuses on the task of identifying the underlying topics being discussed in spoken audio recordings. Both supervised topic classification and topic clustering approaches are discussed in detail.

Chapter 13 Speech Summarization

This chapter focuses on approaches towards automatic summarization of spoken documents, such as meeting recordings or voicemail. While summarization is a well-studied area in natural language processing, its application to speech is relatively recent, and this chapter focuses on extending text-based methods and evaluation metrics to handle spoken input.

Chapter 14 Speech Analytics

This chapter attempts to provide a detailed description of techniques towards speech analytics or speech data mining. Since this task is rooted in commercial applications, especially in call-centers, there is very little published work on the established methods, and in this chapter we aim to fill this gap.

Chapter 15 Speech Retrieval

This chapter discusses the retrieval and browsing of spoken audio documents. This is an area lying between the two distinct scientific communities of information retrieval and speech recognition. This chapter aims to provide an overview of the common tasks and data sets, evaluation metrics, and algorithms most commonly used in this growing area of research.

References

Allen J 1995 *Natural Language Understanding* Benjamin/Cummings, Chapter 8.

Chomsky N 1965 *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Jackendoff R 2002 *Foundations of Language* Oxford University Press, Chapter 9.

Kuhn R and De Mori R 1995 The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17**, 449–460.

Lowe JB and Baker CF 1997 A frame-semantic approach to semantic annotation *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)-SIGLEX Workshop*, Washington, D.C.

Pieraccini R, Tzoukermann E, Gorelov Z, Gauvain JL, Levin E, Lee CH and Wilpon JG 1992 A speech understanding system based on statistical representation of semantics *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, CA.

Price PJ 1990 Evaluation of spoken language systems: The ATIS domain *Proceedings of the DARPA Workshop on Speech and Natural Language*, Hidden Valley, PA.

Seneff S 1992 TINA : A natural language system for spoken language applications. *Computational Linguistics* **18**(1), 61–86.

Shannon CE 1948 A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423, 623–656.

Turing AM 1950 Computing machinery and intelligence. *Mind* **49**(236), 433–460.

Ward W and Issar S 1994 Recent improvements in the CMU spoken language understanding system *Proceedings of the ARPA Human Language Technology Conference (HLT) Workshop*, pp. 213–216.

Weizenbaum J 1966 Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the ACM* **9**(1), 36–45.