Spoken Language Understanding

1

Human/Human Conversation Understanding

Gokhan Tur and Dilek Hakkani-Tür
Speech at Microsoft | Microsoft Research

While the term spoken language understanding mostly refers to the understanding of spoken utterances directed at machines, as they are more constrained, recent progress in recognition and understanding the human/human conversations and multi-party meetings is not negligible. While there is significant amount of previous work on discourse processing especially in social sciences (such as in the field of conversation analysis), processing human/human conversations is a relatively newer area for spoken language processing.

In this chapter we focus on two-party and multi-party human/human conversation understanding approaches, mainly focusing on discourse modeling, speech act modeling, and argument diagramming. We also try to point the bridge studies using human/human conversations for building better human/machine conversational systems or using the approaches for human/machine understanding for better human/human understanding and vice versa.

1.1 Background

While speech is the most natural medium of human/human communication, little of spoken data is available for research purposes due to privacy or copyright issues or signal quality related issues, such as non-ideal recording conditions - even in some call centers. Unlike textual communication, such as emails or instant messaging, almost all of spoken interactions are lost unrecorded, unprocessed.

Arguably the most famous human/human conversations corpus, Switchboard (Godfrey et al. 1992), has been collected in the early 90s, sponsored by the DARPA. It is a large multi-speaker corpus with 2430 telephone conversations of 3 to 10 minutes duration (averaging 6 minutes) each, spoken by about 500 paid volunteers on pre-determined topics, totaling 240 hours of speech and about 3 million words of text. This corpus has first been designed to complement the TIMIT corpus (Zue et al. 1990), which was read speech, and the ATIS corpus (Price 1990), which was human/machine interactions. Figure 1.1 shows an example excerpt from the Switchboard corpus. The original targeted research areas were speech and

Spoken Language Understanding Edited by Gookhan Tur and Renato deMori

- ...
- 172.16 175.05 A: because anytime you can walk through a door make ten dollars an hour. laugh
- 175.20 176.87 B: %oh, that's, that's very good.
- 175.47 176.92 A: i, i don't, i don't see any
- 176.92 179.25 A: ((well)) yes it is very good, you know?
- 178.02 179.31 B: mhm, that's ((really)) good.
- 179.80 182.61 B: i know a friend of mine also, he worked at the post office
- 182.63 186.70 B: and, i worked at the post office a long time ago, ((and)) it was very powerful.
- 182.87 183.35 A: mhm.
- 185.64 186.63 A: ((y)) so did i.
- 187.03 188.97 B: oh you did? laugh okay laugh.
- 187.07 187.48 A: mhm.
- 187.84 188.88 A: yes i did.
- 189.03 189.75 B: {breath} and
- 189.42 192.41 A: as a matter of fact, i worked at the ∼DC office, north ˆ Capital street.
- 192.46 193.38 B: %oh okay.
- 193.46 194.01 A: yeah.
- 193.91 195.49 B: and when, remember when
- 195.55 198.07 B: well, y- i don't know how old you are, but i'm telling my age.
- 198.09 201.89 B: (( )) when the post off- when the postal workers, when they went on strike
- 198.59 198.99 A: {laugh}
- ...

Figure 1.1 An excerpt transcription from the Switchboard corpus with disfluency (%), lower human transcriber confidence (double parentheses) nonspeech (e.g., "breath"), and named entity (∼) annotations.

speaker recognition instead of speech understanding. This corpus has indeed been extensively used in these areas and is still being used. After the success of this study, a follow-up collection, named Switchboard-II, has been performed.

The need for more natural and multilingual/accented conversational speech corpora has then led to collections named as CallHome and Call-Friend, which have small number of speakers making relatively longer telephone conversations to people they already know.

Furthermore, goal-oriented human/human dialogues have been collected, which is a mix of ATIS and Switchboard corpora. Some notable studies include the TRAINS (Allen et al. 1995) corpus for transportation planning, the Monroe (Stent 2000) corpus for disaster handling, and the MapTask (Anderson et al. 1991) corpus for giving directions on a map. TRAINS is a collection of task oriented human/human dialogs for the transportation domain. The speaker tries to ship some boxcars to some location. The other human, playing an assistant system

which has access to extra information tries to help the speaker.

With the start of another large scale DARPA program, named EARS (Effective, Affordable, Reusable Speech-to-Text), a larger corpus, named Fisher, has been collected with 16,454 English conversations, totaling 2,742 hours of speech. This data has been transcribed using LDC "quick transcription" specifications, that include a single pass with some automated preprocessing.

The speech processing community has then studied extensions of two-party human/human conversations in a few directions: multi-party human/human conversations (or meetings), lectures, and broadcast conversations (such as talkshows, broadcast discussions, etc.).

Projects initiated at CMU (Burger et al. 2002) and ICSI (Janin et al. 2004) in the late 1990s and early 2000s collected substantial meeting corpora and investigated many of the standard speech processing tasks on this genre. Subsequently, several large, interdisciplinary, and multi-site government-funded research projects have investigated meetings of various kinds. The AMI (Augmented Multi-party Interaction) Consortium (AMI n.d.) and DARPA-funded CALO (Cognitive Assistant that Learns and Organizes) (*DARPA Cognitive Agent that Learns and Organizes (CALO) Project* n.d.) projects concentrate on conference-room meetings with small numbers of participants. The CHIL (Computers in the Human Interaction Loop) project (CHI n.d.) collected a series of lectures dominated by a single presenter with shorter question/answer portions, as well as some "interactive" lectures involving smaller groups. AMI and CHIL also produced corpora of time-synchronized media, generally including close-talking and far-field microphones, microphone arrays, individual and room-view video cameras, and output from slide projectors and electronic whiteboards.

Starting in 2002, the annual NIST Rich Transcription (RT) Evaluations have become a driving force for research in conversational speech processing technology, with substantial performance improvements in recent years. In order to promote robustness and domain-independence, the NIST evaluations cover several genres and topics, ranging from largely open-ended, interactive chit-chat, to topic-focused project meetings and technical seminars dominated by lecture-style presentations. NIST evaluates only the speech recognition and speaker diarization systems, with a focus on recognition from multiple distant table-top microphones.

However, technology has advanced such that many other types of information can be detected and evaluated — not least including dialog acts, topics, and action items. In the next sections, we will try to cover basic understanding tasks studied in the literature.

1.2 Human-Human Conversation Understanding Tasks

Conversations between humans is the most natural and efficient way of communication. Various types of conversational setups such as two-party conversations, multi-party meetings, and lectures exist in everyday life of people. For example, within organizations, people meet for various reasons, such as discussing issues, task assignments, and planning. While such conversational interactions are so common, there is still no globally adopted automated or semi-automated mechanism for tracking conversations, saving the conversation content for later use by participants or non-participants, or automatically annotating certain content related features such as topics discussed or argued, decisions made.

Recently, the availability of conversational speech corpora, as discussed earlier, and shared task evaluations such as the ones performed by NIST, facilitated research on

automatic processing of conversational speech. Furthermore, in general, such human-human communication in stored audio form has rapidly grown in providing ample source material for later use. In particular, the increased prominence of search as a basic user activity has meant that the ability to automatically browse, summarize, and graphically visualize various aspects of the spoken content has become far more important.

However, there are still open questions: What information from these interactions would be useful for later use? Does this information depend on user/purpose? How could the transcripts of these conversations be used? Our goal in this chapter is not to answer these questions, but provide an overview of some of the research done in the last decade to process human/human conversational speech for providing access to their contents:

- *Dialog act tagging*: aims to annotate speech acts such as suggestions, questions, jointly with conversational linguistics acts such as acknowledgment, agreement, and so on. This task is heavily influenced by annotated human/human and multiparty meetings corpora such as Switchboard and ICSI/AMI data sets. This task is mainly treated as an enabling technology for further conversation processing, such as extracting action items or discussions. Dialog act tagging is generally framed as an utterance classification problem (Stolcke et al. 2000; Tur et al. 2006, among others), following the dialog act segmentation.

- *Dialog act segmentation*: aims to chop a spoken utterances into dialog act units as defined by the dialog act tagging schema followed. While this is a task very close to sentence segmentation, there are certain nuances due to the nature of spontaneous conversational speech. Dialog act segmentation is treated as a binary boundary classification problem using lexical, prosodic, and acoustic features (Kolar et al. 2006; Shriberg et al. 2000, among others). There are also few studies performing joint dialog act segmentation and tagging (Warnke et al. 1997; Zimmermann et al. 2005). We will cover both dialog act tagging and segmentation in Section 1.3 after providing detailed information on various dialog act tagging schema.

- *Discourse and topic segmentation*: aims to chop a conversation into topically coherent units. While topic segmentation of text or prewritten speech is a well established area, there are relatively fewer studies on processing human/human conversations and multiparty meetings. We cover this area in Chapter **??** in this book.

- *Summarization*: aims to generate a compact, summary version of meeting discussions. These summaries can be formed by extracting original speaker utterances (extractive summarization) or by formulating new sentences for the summary (abstractive summarization). Speech summarization is covered in detail in Chapter **??**.

- *Action item and decision detection*: aims to detect task assignments to people and associated deadlines and decision making subdialogs during a meeting. These can be used to enter such information into the related person's calendar, or to track status and progress in the following meetings. The decisions made in meetings can be used for indexing meetings, and one can go back and access the content of the meeting where a specific decision was made. They can also be used to track progress and efficiency of meetings. We will cover some preliminary work towards detecting action items and decisions in formal multiparty meetings in Section 1.4.

- *Agreement/disagreement detection*: aims to mark agreements and disagreements between meeting participants with the goal of supporting meeting summarization, and decision and action item detection. While this task can be seen under dialog act modeling, depending on the following application it will be used for, it can be treated as a special subtask. We will cover this task under dialog act tagging, action item and decision detection sections.
- *Subjectivity and opinion detection*: aims to mark subjective content in meetings, such as opinions, sentiments and arguing. While there is a large literature on text processing in these areas, speech processing is relatively newer but growing very quickly. This topic will be analyzed in Section 1.7.
- *Modeling Dominance*: aims to detect dominant speakers spoken conversations using acoustic and lexical cues. Dominant speakers are defined as trying to assert authority by manipulating the group or certain individuals in the group. Other nonverbal features, such as speaking time, and verbal cues, such as number of words spoken have also been found to be useful. This new area is briefly covered in Section 1.9.
- *Speaker role detection*: aims to classify each of the speakers with respect to their institutional roles. While this is an area deeply rooted in social sciences, most systems have taken a simplistic view and instead focused on professional roles, such as professor vs. student, boss vs. employee, or project manager vs. software engineer. This topic will be covered in Section 1.8.
- *Hot spot detection*: aims to mark regions in which participants are highly involved in the discussion (e.g., heated arguments, points of excitement, and so on). This relatively newer area in processing conversations will be covered in Section 1.6.
- *Addressee detection*: An important enabling task in processing conversations is to determine who is talking or referring to who. This is useful in downstream understanding of conversations in that it provides essential semantic grounding to the analyzed dialog. This task also covers resolving the (especially pronominal) references in utterances. More detailed information is provided in Section 1.5.
- *Argument diagramming*: aims to display the flow and structure of reasoning in conversations, especially in discussions and arguments. For example, one meeting utterance may open a new issue and another utterance may elaborate on it in response. While deeply rooted in social sciences, there is an emerging interest to this task and some preliminary studies. Argument diagramming is described in more detail in Section 1.10.

1.3 Dialog Act Segmentation and Tagging

A speech act is a primitive abstraction or an approximate representation of the (typically) illocutionary (rather than locutionary or perlocutionary) force of an utterance, such as asking, answering, promising, suggesting, warning, or requesting. The communicative speech act theory goes back to the 1960s, when Austin (Austin 1962) defined an utterance in a conversation as a kind of action being performed by the speaker. Later Searle modified the taxonomy of Austin into five major classes (Jurafsky and Martin 2008):

- Assertives (or representatives) for committing the speaker to something is being the case such as suggesting, concluding
- Directives for attempts by the speaker to do something such as ordering or advising
- Commissives for committing the speaker to some future action such as planning or betting
- Expressives for expressing the psychological state of the speaker such as thanking or apologizing
- Declarations for bringing about a different state of the world such as "I name this ship the Titanic"

While this taxomony covers many of the actions conveyed in speech, it ignored the *conversational* aspect of the spoken interactions such as grounding and contributions such as acknowledgments or backchannels. According to Clark and Schaefer, a conversation is a series of presentations and acceptances (Clark and Schaefer 1989). In the first phase, a speaker performs a kind of speech act as explained above. The difference is in the second phase where the hearer acts upon this speech act. In this acceptance phase, the hearer *grounds* the previous utterance in a variety of methods such as acknowledging, repeating or paraphrasing the previous utterance, utters on the next relevant contribution, or simply shows continued attention:

A: I worked at the DC office as the branch manager

- B1: Oh, okay (acknowledgment)
- B2: And that was in 2000, right? (next contribution)
- B3: Oh, you were the manager there (paraphrasing)

This more complex framework of joint linguistics acts with two phases is studied under the name of *dialog act tagging*. The main goal of dialog acts is to provide a basis for further discourse analysis and understanding. For example, it has been shown that dialog acts can be used to extract the action items or question/answer pairs in conversations as discussed later. Furthermore, as dialog acts are designed to be task independent, it is easier to reuse them or use them as a starting point when processing new genre.

1.3.1 Annotation Schema

There are a number of dialog act tagging schema proposed in the literature. Two popular contemporary dialog act sets for conversational speech in the literature are DAMSL (Dialog Act Markup in Several Layers (Core and Allen 1997) and MRDA (Meeting Recorder Dialog Act) (Shriberg et al. 2004b). Next we will cover these two and others in more detail.

DAMSL focuses on providing multiple layers of dialog act markup. Each layer allows multiple communicative functions of an utterance to be labeled. For example an utterance can simultaneously perform actions such as responding a question, confirming understanding, promising to perform an action, and informing. DAMSL is organized into three main categories, one for the speech acts, another for the acceptance phase, and a third set for covering extra dimensions of communication:

- The Forward Communicative Functions consist of a taxonomy in a similar style as the actions of traditional speech act theory, namely representatives (or statements), directives, commissives, and performatives (or declaratives). Since it has been designed for task-oriented dialogs, they added one more category, called Open-Option, where a speaker gives a potential course of action but does not show preference toward it (hence different than suggestion) such as in "how about this?".
- The Backward Communicative Functions indicate how the current utterance relates to the previous dialog, corresponding to the acceptance phase above, such as accepting a proposal confirming understanding or answering a question. It consists of four categories: Agreement classes, namely Accept, Reject, Maybe, Hold, Partial-Accept and Partial-Reject, understanding classes, namely Acknowledgment, Repeat/Paraphrase, Completion, Correct misspeaking, answering. They also added one class for signaling non-understanding, such as "huh?" or "you mean, to Dansville?", and another class for covering utterances with an information relation to the previous one, such as providing examples, elaborating on it, etc.
- The Utterance Features include information about an utterances form and content such as whether an utterance concerns the communication process itself or deals with the subject at hand. It has three subcategories: The information level identifies whether an utterance is about a task, task management (such as "What times are available"), or else (communication management) utterances. The communicative status is for the abandoned and uninterpretable utterances. The syntactic features (a misnomer) cover generic communicative utterances such as "hello", and exclamations such as emotional utterances.

Figure 1.2 provides a summary of the three main categories of annotations. Some popular corpora annotated with DAMSL tags are the TRAINS (Allen et al. 1995) corpus for transportation planning, the Monroe (Stent 2000) corpus for disaster handling, the MapTask (Anderson et al. 1991) corpus for giving directions on a map, and the Switchboard (Godfrey et al. 1992) corpus.

Jurafsky et al. (1997) have adopted the DAMSL schema for the Switchboard corpus. They managed to exploit about 80% of the DAMSL tags, but for certain cases they made some changes. For example, they added non-verbal (such as breath) and third party talk to communicative status, formed subcategories for answer (such as yes/no answer, descriptive answer, etc.), and similarly for the information request set (such as yes/no question, open question, etc.), merged assert and re-assert, and marked hedge. They ended up with about 60 tags, which can be clustered into 40 categories.

Note that, not all dialog act tagging schema creation efforts have been motivated by the existing social theories. A notable example is the Verbmobil dialog act tagging schema, which is used for planning schedules (Susanne et al. 1995). The designers have come up with a tagset motivated by the existing corpus instead of the other way around. They have generic tags such as thanks, greet, introduce, by, a subset of traditional speech acts such as request or suggest, and a subset of acceptance tags such as reject, accept, confirm, clarify, and a number of task related tags such as backchannel, garbage, give-reason (e.g., "because I have meetings all afternoon"), and initial-request (e.g., "I want to make an appointment for tomorrow"). At this point, also note the relationship between the dialog act tagging and the intent determination task studied for human/machine interactions covered in Chapter **??**.

Forward Communicative Functions (Speech act phase)

- Statements (Assertives or Representatives)
- Directives and Open-Option
- Commitments
- Performatives
- Other

Backward Communicative Functions (Acceptance phase)

- Agreements
- Understandings
- Answers
- Information-Relation

Utterance Features

- Information Level
- Communicative Status
- Syntactic Features

Figure 1.2 The DAMSL annotation categories.

A different view for conversation act tagging has been proposed by Traum and Hinkelman (1992). Their schema consists of four layers depending on the unit to be tagged:

- Turn taking tags for sub-utterance units. These are mainly floor mechanisms for keeping or releasing the turn.
- Speech acts for discourse units. These include traditional speech acts such as suggest, request, questions, etc. But instead of tagging one utterance, a set of utterances called discourse units are tagged. A discourse unit consists of an initial presentation and subsequent utterances until act is mutually understood (or grounded). Hence, the shortest discourse unit is an initial presentation and an agreement or acknowledgment.
- Grounding tags for utterance units. These correspond to single utterances or sentential units for the acceptance phase, such as for acknowledgment, continue, or repair.
- Argumentation acts for one or more discourse units. In the most basic form, a question/answer pair is an argumentation act with one discourse unit. However, argumentation acts can be built up hierarchically. For example, a typical conversation from the TRAINS corpus can start with goal specification, followed by planning and then verification of the plan. These tags are also very closely related to the argument diagramming of the conversations which we will cover later.

The latter popular dialog act tag annotation scheme, MRDA, focuses on multi-party meetings. This schema has been used to annotate the ICSI meeting corpus, a collection

Tag	DAMSL	MRDA	Tag	DAMSL	MRDA
Indecipherable	%	%	Conventional Opening	fp	
Abandoned	%-	%–	Conventional Closing	fc	
Interruption		%-	Topic Change		tc
Nonspeech	x	x	Explicit-Performative	fx	
Self-Talk	t1	t1	Exclamation	fe	fe
3^{rd}-Party Talk	t3	t3	Other Forward Function	fo	
Task Management	t	t	Thanks	ft	ft
Communication Management	c		Welcome	fw	fw
Statement	sd	s	Apology	fa	fa
Subjective Statement	sv	s	Floor Holder		fh
Wh- Question	qw	qw	Floor Grabber		fg
Y/N Question	qy	qy	Accept, Yes	ny, aa	aa
Open Ended Question	qo	qo	Partial Accept	aap	aap
Or Question	qr	qr	Partial Reject	arp	arp
Or Clause After Y/N Question	qrr	qrr	Maybe	am	am
Rhetorical Question	qh	qh	Reject, No	nn, ar	ar
Declarative Question	d	d	Hold	h	h
Tag Question	g	g	Collaborative Completion	2	2
Open Option	oo		Backchannel	b	b
Command	ad	co	Acknowledgment	bk	bk
Suggestion	co	cs	Mimic	m	m
Commit	cc	cc	Repeat		r
Reformulation	bf	bs	Appreciateion	ba	ba
Sympathy	by	by	Downplayer	bd	bd
Misspeak Correction	bc	bc	Nonlabeled		z
Rhetorical-Question Backchannel	bh	bh	Signal Non-understanding	br	br
Understanding Check		bu	Defending/Explanation		df
Misspeak Self-Correction		bsc	”Follow me”		f
Expansion/Supporting Addition	e	e	Narrative-Affirmative Answers	na	na
Narrative-Negative Answers	ng	ng	No-Knowledge Answers	no	no
Dispreferred Answersnd	nd	nd	Quoted Material	q	
Humorous Material		j	Continued From Previous Line	+	
Hedge	h				

Figure 1.3 SWBD-DAMSL and MRDA tag sets. Italic means slightly modified meaning.

of 75 about one hour long multiparty dialogs, naturally occurring in an academic setting. While it is similar to SWBD-DAMSL, one big difference is that it includes a set of labels for floor management mechanisms, such as *floor grabbing* and *holding*, which are common in meetings. Furthermore additional tags were used for topic changes and humorous material such as jokes. Interrupted sentences were also assigned their dedicated tags, different than abandoned sentences. Statements and subjective statements are combined as well. The tag sets for MRDA and DAMSL extended for the Switchboard corpus (SWBD-DAMSL) schema are presented in Figure 1.3. A sample meeting (actually about the meetings project itself) is also presented in Figure 1.4.

Clark and Popescu-Belis has also proposed a more shallow tagging schema called MALTUS, which is easier to tag automatically (Clark and Popescu-Belis 2004). MALTUS basically clustered some of the MRDA tags and dropped some others. An utterance is then tagged with one of the four high level tags: statement, question, backchannel, and floor

Time	Speaker	DA Tag	Transcript
2804-2810	c3	s ^df ^e.%-	I mean you can't just print the values in ascii ==
2810-2811	c6	fg	well ===
2810-2811	c5	s ^arp ^j	not unless you had a lot of time
2811-2812	c5	%-	and ==
2811-2814	c6	s ^bu	uh and also they're not - I mean as I understand it you - you don't have a way to optimize the features
2814-2817	c6	qy ^d ^g ^rt	right?
2818-2818	c2	s ^aa	right

Figure 1.4 A sample excerpt from the ICSI meeting corpus annotated with MRDA tags. ^ is used for multiple labels for the same dialog act unit and | is used for consecutive dialog act tags (from (Shriberg et al. 2004b)).

mechanism, followed by one of the eight subcategories of them, which are attention, action, correction/repetition, politeness, and positive/negative/undecided responses. More detailed information about various dialog act schema and comparisons can be found in (Popescu-Belis 2005).

1.3.2 Modeling Dialog Act Tagging

Dialog act tagging is generally framed as an utterance classification problem (Mast et al. 1996; Stolcke et al. 2000; Tur et al. 2006, among others). Previous studies differ in the features and classification algorithms they have employed and the dialog act set (as presented before) they use. The range of features include lexical (such as word n-grams), prosodic (such as pitch and energy), syntactic, contextual (such as the previous and next estimated dialog act tags), and others (such as number of words, duration), optionally normalized by the speaker or the conversation.

The basic approach as taken by Tur et al. (2006) is to treat each sentence independently and to employ lexical features, that is word n-grams and number of words in the sentence in classifiers. They used the ICSI meeting corpus with high-level MRDA tags: *question*, *statement*, *backchannel*, *disruptions*, and *floor grabbers/holders* with a Boosting classifier.

Mast et al. (1996) used semantic classification trees, similar to the approach taken by Kuhn and Mori (1995). They are basically decision trees where the nodes query about existence of phrases. Following the decision tree training literature back then, they used the Gini criterion to decide on convergence. They showed performance results using 19 dialog act tags on the German VerbMobil spontaneous speech corpus.

The first study using prosodic information for dialog act tagging was presented in (Shriberg et al. n.d.). They used decision trees to query also prosodic features such as pitch, duration, pause, energy, and speaking rate. They demonstrated that prosody contributed significantly to improving dialog act tagging performance both when using manual and automatic speech transcriptions of the Switchboard corpus, following the SWBD-DAMSL schema, as explained above.

Stolcke et al. (2000) presented a more comprehensive system for classifying dialog acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of

the dialog act sequence. The dialog model is based on treating the discourse structure of a conversation as an HMM and the individual dialog acts as observations emanating from the model states. Constraints on the likely sequence of dialog acts are modeled via a dialog act n-gram. The statistical dialog act grammar is combined with word n-grams, decision trees, and neural networks modeling the idiosyncratic lexical and prosodic manifestations of each dialog act. They also reported performance figures using the Switchboard spontaneous speech corpus.

Following these studies, Venkataraman et al. tried employing active learning and lightly supervised learning for reducing the amount of labeled data needed for dialog act tagging with HMMs. They concluded that while active learning does not help significantly for this task, exploiting unlabeled data by using minimal supervision is effective when the dialog act tag sequence is also modeled (Venkataraman et al. 2005, 2002).

Tur et al. (2006) proposed model adaptation methods for dialog act tagging. They used the ICSI meeting corpus with five high-level meeting recognition dialog act tags, and performed controlled adaptation experiments using the Switchboard (SWBD) corpus with SWBD-DAMSL tags as the out-of-domain corpus. They obtained significantly better performance by automatically selecting a subset of the Switchboard corpus and combining the confidences obtained by both in-domain and out-of-domain models via logistic regression.

Margolis et al. (2010) presented an extension of this study recently, focusing on cross-lingual adaptation using non-lexical features and using machine translation output in the other language. They have performed experiments using MRDA and SWBD, similar to (Tur et al. 2006) for English and the Spanish Callhome corpus for cross-lingual adaptation. They have mainly explored the use of structural correspondence learning (SCL) for domain adaptation, which relies on feature transformation of learned pivot features, which are informative for both domains.

More recently, Rangarajan et al. (2007) proposed exploiting prosodic and syntactic features in a Maximum Entropy classification framework. Their contribution was using a syntax-based categorical prosody prediction. They reported improved results using the Switchboard corpus over the baseline using only lexical features.

1.3.3 Dialog Act Segmentation

Dialog act segmentation is a crucial first step in processing conversational speech such as meetings or broadcast conversations as most of the follow-up processing such as summarization or argument diagramming rely on sentential units. Dialog act segmentation is generally framed as a word boundary classification problem. More formally, the goal is finding the most likely boundary tag sequence, T given the features, F, in a sentence:.

$$argmax_T P(T|F)$$

To this end mostly generative, discriminative, or hybrid models are used. The most well known generative model is the hidden event language model, as introduced by (Stolcke and Shriberg 1996). In this approach, sentence boundaries are treated as the hidden events and the above optimization is simply done by the Viterbi algorithm using only lexical features, i.e. language model.

For example, Figure 1.5 shows the model for the two-class problem: *nonboundary* (NB) and *sentence boundary* (SB) for sentence segmentation. Table 1.6 shows an example

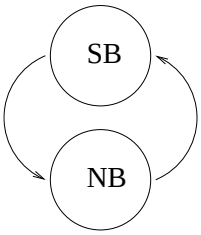

Figure 1.5 Conceptual hidden Markov model for segmentation with two states: one for segment boundaries, one for others.

Emitted Words	...	people	are	dead	few	pictures	...
State Sequence	...	NB	NB	SB	NB	NB	...

Figure 1.6 Sentence segmentation with simple 2-state Markov model.

sequence of words emitted. This method was extended with confusion networks in (Hillard et al. 2004).

Note that this is not different from using an HMM as is typically done in similar tagging tasks, such as part of speech (POS) tagging (Church 1988) or named entity extraction (Bikel et al. 1999). However, it has been shown that the conventional HMM approach has certain weaknesses. For example, it is not possible to use any information beyond words, such as POS tags of the words or prosodic cues for speech segmentation.

To this end, two simple extensions have been proposed: Shriberg et al. (2000) suggested using explicit states to emit the boundary tokens, hence incorporating nonlexical information via combination with other models. This approach is used for sentence segmentation and is inspired by the hidden event language model (HELM), as introduced by (Stolcke and Shriberg 1996), which was originally designed for speech disfluencies. The approach was to treat such events as extra meta-tokens. In this model, one state is reserved for each boundary token, SB and NB, and the rest of the states are for generating words. To ease the computation, an imaginary token is inserted between all consecutive words, in case the word preceding the boundary is not part of a disfluency. The following example is a conceptual representation of a sequence with boundary tokens:

... people NB are NB dead SB few NB pictures ...

The most probable boundary token sequence is again obtained simply by Viterbi decoding. The conceptual HELM for segmentation is depicted in Figure 1.7.

These extra boundary tokens are then used to capture other meta-information. The most commonly used meta-information is the feedback obtained from other classifiers. Typically, the posterior probability of being in that boundary state is used as a state observation likelihood after being divided by prior probabilities. These other classifiers may be trained with other feature sets as well, such as prosodic or syntactic, using decision trees to get hybrid models (Shriberg et al. 2000).

The second extension is inspired from factored language models (fLMs) (Bilmes and Kirchhoff 2003), which capture not only words but also morphological, syntactic, and

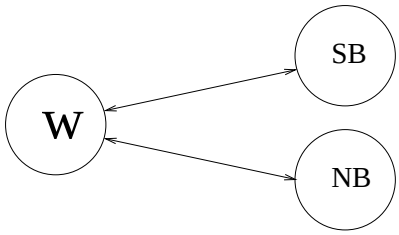

Figure 1.7 Conceptual hidden event language model for segmentation.

other information. (Guz et al. 2009) proposed using factored HELM (fHELM) for sentence segmentation using POS tags in addition to words for incorporating morphological information, which is especially important for inflectional languages.

With the advances in discriminative classification algorithms, researchers tried using Conditional Random Fields (Liu et al. 2005), Boosting (Cuendet et al. 2006), multi-layer perceptrons (Mast et al. 1996), and hybrid approaches using Boosting and Maximum Entropy classification algorithms (Zimmerman et al. 2006). Features can be the presence of specific word n-grams around the candidate boundary, an indicator of being inside a quotation in text, an indicator of presence of the preceding word tokens in an abbreviation list, or duration of pause, pitch, energy, and other duration-related features in speech.

Zimmerman et al. (2006) provides an overview of different classification algorithms (Boosting, hidden-event language model, maximum entropy and decision trees) applied to the dialog act segmentation for multilingual broadcast news. They concluded that hybrid approaches are always superior and (Guz et al. 2009) concluded that this is also true with CRF, although to a lesser degree.

So far, most approaches to sentence segmentation have focused on recognizing boundaries rather than sentences in themselves. This has occurred because of the quadratic number of sentence hypotheses that must be assessed in comparison to the number of boundaries. To tackle that problem, Roark et al. (2006) segment the input according to likely sentence boundaries established by a local model, and then train a reranker on the n-best lists of segmentations. This approach allows leveraging of sentence-level features such as scores from a syntactic parser or global prosodic features. Favre et al. (2008b) proposed to extend this concept to a pruned sentence lattice, which allows combining of local scores with sentence-level scores in a more efficient manner.

Recent research has focused on model adaptation methods for improving dialog act segmentation for meetings using spontaneous telephone conversations, and speaker-specific prosodic (Kolar et al. 2007) and lexical modeling (Cuendet et al. 2006). Guz et al. (2010) has studied the effect of model adaptation of segmentation and tagging jointly applied in a cascaded manner. There are also a number of studies studying the impact of segmentation on follow-up tasks such as summarization (Liu and Xie 2008), information extraction (Favre et al. 2008a), and translation Matusov et al. (2007).

1.3.4 Joint Modeling of Dialog Act Segmentation and Tagging

While most systems first segment input speech utterances into dialog act units and then assign dialog act tags (e.g., (Ang et al. 2005; Mast et al. 1996; Tur et al. 2010, among others)), there are a number of studies proposing joint segmentation and tagging. Two notable ones include the following: Warnke et al. (1997) proposed to use the A* algorithm to search for optimal segmentation and classification of dialog acts. The candidates for boundaries are determined using a multi-layer perceptron classifier and the dialog act candidate tags are determined using statistical language models. Then the task of the integration is choosing the optimal joint segmentation and tagging. Zimmermann et al. (2005), on the other hand, proposed a totally joint approach, without any underlying segmentation and tagging components. Their approach is based on generative approaches for segmentation, and instead of using only two boundary classes (e.g., SB and NB in the above example), they reserved one boundary class per dialog act tag. Note that, using hidden event models, this approach may be used to build hybrid joint classification models trained with prosodic or other features.

1.4 Action Item and Decision Detection

A practical high level task for understanding human/human conversations is extracting key information related to action items and decisions. These tasks are also among the most commonly requested outputs from meetings according to user studies (Banerjee et al. 2005; Lisowska 2003).

However, it is not clear what constitutes a decision or an action item. For example, one would argue that an action item consists of an assertion, command, or suggestion (e.g., "do this until tomorrow") which has optionally an assignee and a due date, followed by an agreement. However it is not hard to think cases where such a simplistic view fails to cover what is intended in this task (e.g., "open the window now"), as natural language is nothing but gray areas when it comes to semantic processing.

While some researchers have treated this task as a subtask of argument diagramming (Section 1.10), putting it in a larger scope (e.g., (Verbree et al. 2006)), a number of studies have directly attacked these tasks after carefully limiting the scope of the task and the conversation style to minimize these gray areas (more formally, to increase the interannotator agreement). For example, consider the excerpt in Figure 1.8 from the CALO corpus, where there is a clear task that needs to be done and an assignee.

More specifically, both tasks are defined as marking dialog act units which have information about action items (typically the task itself together with the due date and responsible party) and decisions (the issue involved and the resolved course of action).

There is some related work in the text processing domain, focusing on the emails under the DARPA PAL program, using Enron email corpus[1]. This is a binary classification task, trying to detect whether an email (or a sentence in an email) has any content related to any action item (Bennett and Carbonell 2005, 2007; Corston-Oliver et al. 2004). F-scores around 80% are achieved on the task of classifying emails as containing action items, and 60% to 70% when classifying individual sentences. They used word n-grams with SVM classifiers.

To process spoken multi-party conversations, similar approaches were employed but with mixed results. Some success has been shown in detecting decision-making utterances in

[1] http://www.cs.cmu.edu/ enron/

- *John Smith*: so we need to arrange an office for joe browning (statement/all)
- *Kathy Brown*: are there special requirements (question/John)
- *Cindy Green*: when is he co- (disruption/John)
- *John Smith*: yes (affirmation/Kathy) // there are (statement/Kathy)
- *John Smith*: we want him to be close to you (statement/Kathy)
- *Kathy Brown*: okay (agreement/John) // I'll talk to the secretary (commitment/John)
- *Cindy Green*: hold on (floor grabber/all) // wh- when is he coming (question/John)
- *John Smith*: next monday (statement/Cindy)
- *Cindy Green*: uh-huh (backchannel/all)

Action Item: Arrangement of Joe's office location
Owner: Kathy

Decision: Location of Joe's office to be close to Kathy

Summary:

- *John Smith*: so we need to arrange an office for joe browning (statement/all)
- *John Smith*: we want him to be close to you (statement/Kathy)

Figure 1.8 An excerpt from the CALO meeting corpus. Dialog act tags and addressed persons are shown in parentheses. This meeting data has one action item and one decision. A brief extractive summary corresponding to this meeting data follows.

meetings in a constrained domain. For example, Hsueh and Moore (2007) achieve F-scores of 60% to 70% for the task of detecting decision-making utterances as identified from within a manually selected summary set using the AMI corpus. They employed a Maximum Entropy classifier using lexical (words and phrases), prosodic (pitch and intensity), semantic (dialog act tags, temporal expressions) and contextual (relative position within the meeting) features. This study assumed that they have the manual transcriptions and annotation of dialog act tags, topic boundaries and topics.

On the other hand, when the task is to detect utterances from within an entire meeting, and when the domain is less constrained, accuracy seems to suffer significantly: Morgan et al. (2006) achieved F-scores of only around 30% when detecting action item utterances over the ICSI Meeting Corpus using similar features.

This shows the importance of the meeting style for this task: structured and artificial project meetings of the AMI corpus versus unstructured, naturally occurring meetings of the ICSI corpus. In naturally occurring speech, tasks are defined incrementally, and commitment to them is established through interaction between the people concerned; cues to their detection can therefore lie as much in the discourse structure itself as in the content of its constituent sentences.

To this end, for the DARPA CALO project, Purver et al. (2007) and Fernández et al. (2008) took a structural approach to action item and decision detection: utterances are

first classified according to their role in the commitment process (e.g., task definition, agreement, acceptance of responsibility, issue under discussion, decision made) using binary SVM classifiers, one for each possible utterance role, and then action item or decision discussions are detected from patterns of these roles using a binary classifier or a probabilistic graphical model. Four types of action item related utterances are identified: task description, responsible party identification, deadline assignment, and agreement/disagreement of the responsible party. For decisions, they used three categories: utterances which initiate a discussion by raising an issue, utterances which propose a resolution for the raised issue (such as proposals considered or restatements), and utterances which express agreement for a proposed resolution

On manual transcripts of the AMI meeting corpus, the detectors achieve F-scores of around 45% for action items (Purver et al. 2007) and 60% for decisions (Fernández et al. 2008). This is a significant improvement over the baseline results obtained with non-structured detectors trained on the same data, which achieve 37% and 50% F-scores, respectively. When ASR output is used, there is a drop in the detection performance, but this is still above the chance baseline.

They also studied towards building a real-time detector, where the detector runs at regular and frequent intervals during the meeting. It reprocesses recent utterances in case a decision discussion straddles these and brand new utterances, and it merges overlapping hypothesized decision discussions, and removes duplicates. The real-time decision detector does not perform significantly worse than the offline version (Frampton et al. 2009b).

Yang et al. (2008) have explored the use of prosodic cues for improving action item agreement detection. The motivation is that, most agreement phrases are also used for backchanneling (e.g., "yeah"), acknowledgment (e.g., "okay"), or even questions (e.g., "right?"). While the dialog context is critical to disambiguate the correct dialog act tag for them, prosodic features are shown to be helpful, improving the baseline by around 25% for single word utterances.

Related to detecting agreement and disagreements, there are other previous studies which have demonstrated the feasibility of this task, using textual, durational, contextual and acoustic features, on the annotations of the conversational corpora. For example, Hillard et al. (2003) built an agreement/disagreement classifier using decision trees with lexical and prosodic features, on the ICSI meeting corpus. Prosodic features do not improve the performance, but when only the prosodic features are used, the performance degradation due to ASR output was much less, compared to using manual transcriptions, as expected. One thing to note is that, this is a four-way classification task, with positive, negative, neutral utterances and backchannels. Frequent single-word utterances, such as "yeah" or "right" are not considered to be agreements but instead backchannels as they usually reflect encouragement for the speaker to continue more than actual agreement, an assumption which may not be the case all the time.

Galley et al. (2004) exploited adjacency pair information as obtained from addressee detection task as an additional feature while identifying agreement and disagreements. They reported better performance than Hillard et al. using the same corpus with a Maximum Entropy classifier using manual transcriptions. However, most of the improvement comes from the classification algorithm (3% of 4% improvement in accuracy). Their approach can be considered as a preliminary task for argument diagramming based studies (Section 1.10) for detecting agreement/disagreements.

1.5 Addressee Detection and Co-Reference Resolution

In spoken interactions, one particular key task is going beyond automatic transcription and dialog act tagging for each of the utterances and determining the speaker this utterance is addressed to or the entities and individuals referred in the utterance. This is a critical enabling task for downstream conversation understanding tasks such as action item or decision detection or argument diagramming. For example, in order to fully understand the owner of an action item, the system must resolve who this action item is addressed to.

While this task resembles the well-known coreference resolution and mention detection tasks in information extraction, spoken conversations are done jointly between the participants. Speakers and addressees (and overhearers too) are continually trying to establish common ground with each other, using a number of means in different modalities. Eye gaze, physical orientation, backchanneling ("uh-huh", "mm-hmm"), and contextual reference are core components to this process. In contrast to two-party human-computer or human-human dialogues, however, the multi-party nature of meetings presents novel challenges to the analysis of these features. The potential referents and addressees are more numerous, and the complexity of physical interaction increases.

Consider the example in Figure 1.8. Even in this short excerpt there are a number of pronominal references, and especially the resolution of "you" is critical. In a command such as "I want you to do this until this time", it is not necessarily the next speaker who is referred by "you".

The most useful feature for detecting the addressee found in the literature is not the lexical, visual, or contextual cues, but a smart fusion of multimodal features, such as head pose estimation from omnidirectional table-top video, prosodic speech waveform analysis, context, lexical cues, named entities, and higher-level linguistic analysis such as parsing. Stiefelhagen et al. (2002) focused on using visual cues and estimated visual focus of attention at each frame of three 4-person meetings. Using automatically estimated head poses as input to neural networks, they obtained an 8% error reduction (from 74% to 76% accuracy) by adding speaker information.

Research in automatic addressee detection has produced similar results. The task is typically approached as an utterance-level classification where some subset of the participants in the meeting are identified as addressees. Jovanovic et al. (2006) used a combination of lexical features of the utterance (e.g., personal, possessive, and indefinite pronouns, and participant names) and gaze features from each participant to detect addressee in 4-person meetings using Bayesian networks. Multimodality again emerged as a key, with utterance features alone achieving 53% accuracy, speaker gaze 62%, everyone's gaze 66%, and their combination 71%.

In the DARPA CALO meeting assistant project (Tur et al. 2010), without using video, Gupta et al. (2007) automatically detected addressee with 47% accuracy over a 37% baseline by exploring a number of structural, durational, and lexical features taken from the speech transcript only. To leverage dialogue context as well, they used a conditional random field (CRF) as a classifier, looking both forward and backward into the dialogue. For the same project, for processing pronominal references, they employed deeper linguistic analysis. As a pre-processor to a downstream reference resolution system, Müller (2006) used a rule induction classifier to determine whether "it" was referential in meetings from the ICSI corpus. This is important since "it" has also non-referential senses as in "it's raining". 'Note

that, this is also the case for you as in "you know" or "this is how you do that". Furthermore, "you" has singular and plural senses, which makes the addressee detection task much harder. Using several automatically extracted syntactic, lexical, and word distance input features, he achieved an F-score of 62%. Gupta et al. (2007) performed a similar experiment to automatically resolve the referentiality of second-person pronouns, achieving a 63% accuracy over a 58% baseline. Müller (2007) and Gupta et al. (2007) also describe classifiers for resolving these pronouns, if they are indeed classified as referential.

The details of the addressee approach of CSLI used for CALO is presented in (Frampton et al. 2009a; Purver et al. 2009). After distinguishing between generic vs. referential *you* and referential singular versus plurals they identify the individual addressee for the referential singulars. They used Bayesian Networks, using linguistic and visual features, with the AMI meeting corpus. Using a test set composed of around 1000 utterances which contain the word *you*, they computed visual features to indicate at which target each participant's gaze was directed the longest during different periods of time. A further feature indicated with whom the speaker spent most time sharing a mutual gaze over the utterance as a whole. Other features include structural, durational, lexical and shallow syntactic patterns of the *you*-utterance. They also used Backward Looking (BL)/Forward Looking (FL) features, which express the similarity or distance (e.g., ratio of common words, time separation) between the *you*-utterance and the previous/next utterance by each non-speaker. The BL/FL speaker order and the number of speakers in the previous/next 5 utterances are addded as contextual features. Finally, for the manual systems, they also used the AMI dialogue acts of the *you*-utterances, and of the BL/FL utterances. They found out that after visual cues, contextual cues are the most informative.

1.6 HotSpot Detection

Recent interest in the automatic processing of meetings is motivated by a desire to summarize, browse, and retrieve important information from lengthy archives of spoken data. One of the most useful capabilities such a technology could provide is a way for users to locate "hot spots", or regions in which participants are highly involved in the discussion (e.g., heated arguments, points of excitement, and so on). While a subjective task, analyses in (Wrede and Shriberg 2003) found that after training, human raters show good agreement in labeling utterances for involvement. Such regions are likely to contain important information for users who are browsing a meeting for applications of information retrieval.

To enable research, the dialog-act labeled ICSI meeting corpus (Dhillon et al. 2004; Shriberg et al. 2004a) is annotated for hot spots as described in (Wrede et al. 2005). Each hot spot consists of a sequence of one or more dialog acts, by one more speakers. Hot spots have an internal structure, and are also labeled for type. Structural points, such as the hotspot "trigger" or "closure" are determined based on semantics, with reference to hot spot "peaks" in normalized speaker involvement occurring within the hot spot. Type (e.g., disagreement, amusement) are marked, as is the level of "hotness," or perceived speaker affect.

Research has found correlations between types of hot spots and specific dialog acts, as well as between factors such as utterance length and utterance perplexity (Wrede and Shriberg 2003). Certain hot spot types, namely jokes, are associated with higher rates of out-of-vocabulary words for an automatic speech recognizer. Contextual factors such as the individual speaker and meeting type also correlate with overall hot spot production.

As one might expect, increased speaker involvement tends to occur in regions of speaker overlap; however, this association is of only moderate degree. A large-scale study of speaker overlap (Cetin and Shriberg 2006), found that hot spots are about 50% more probable during overlapped speech than overall, but that a 16% overall overlap rate for the meeting corpus increases to only 25% when conditioned on hot spots. Thus, while there is an association between hot spots and overlap, they appear to reflect distinct phenomena.

Automatic detection of hot spots can make use of a multitude of cues. In addition to the factors mentioned above, hot spots show marked acoustic-prosodic features such as pitch and energy. One advantage of such features is that they do not require the output of a speech recognizer. Early work reported in (Wrede and Shriberg 2003) found significant cues to hot spots from speaker-normalized prosodic features, particularly from maximum and mean normalized pitch features, but this is clearly an area for further work.

1.7 Subjectivity, Sentiment, and Opinion Detection

Opinion detection is a very well studied area in natural language processing, especially for processing reviews of certain entities (such as restaurants) or events (such as presidential elections) (Pang and Lee 2008). Application of this field to speech is more recent. Note that opinion or sentiment detection is different than emotion detection, though they are closely related. For example a frustrated or angry person may leave a lousy review for a hotel or restaurant. However, typically they are modeled using very different techniques, for example while emotion detection is typically modeled using mostly audio features, opinion detection mostly relies on lexical features. In this section we focus on opinion detection. For a more comprehensive survey of emotion detection, one may refer to Schuller et al. (2010).

When talking about opinion detection from speech, it is useful to separate monologues used for reviews from interactive conversations. While the first one may seem like a natural extension of review processing, using ASR output instead of text, there is only one study presenting such a system to the best of our knowledge: Polifroni et al. (2010) reported some performance figures from a preliminary experiment in processing audio restaurant reviews using an existing multi-aspect opinion processing framework. Note that this is different than a directed dialog to enter your review but instead focusing on classifying the spoken review according to various aspects related to the restaurant domain, such as food, price, ambiance, etc.

Regarding processing interactive conversations and multi-party meetings, the literature is richer. For example Somasundaran et al. (2006) studied manual annotation of sentiment and arguing using the AMI meeting corpus, using textual and audio information, and showed that human annotators achieve higher annotation agreement when using both textual and audio information, especially for negative sentiment, as shown in Table 1.9, suggesting that automatic systems should also benefit from access to acoustic and prosodic features. In a follow-up study, Wilson (2008) showed similar figures using the same corpus for subjectivity. She proposed having the dimension of subjective vs. objective for positive and negative sentiments. For the sentence “this camera is really bad”, this is subjectively negative, but for the sentence “this camera broke one week after I bought it” is an objective negative sentiment.

Later, Somasundaran et al. (2007) proposed a system for detecting arguing and sentiments in meetings. They used not only word n-grams but also existing sentiment lexicons, dialog

Agreement (Kappa)	Raw Text Only	Raw Text + Speech
Positive Sentiment	0.57	0.69
Negative Sentiment	0.41	0.61

Figure 1.9 Inter-annotator agreement for sentiment annotation using speech and text (from (Somasundaran et al. 2006)).

act tags, and adjacency pairs. The sentiment lexicons, which include positive and negative words, subjectivity indicator phrases, valence shifters, and intensifiers, helped the most on top of the baseline trained using only word n-grams.

1.8 Speaker Role Detection

Social scientists have long understood that the link between the identity of the speakers and the interaction is fundamental, and have established different approaches to explain how participants' embodiment of different identities is relevant for actions in interactions. The most well-known study is about institutional roles by Drew and Heritage (1992) represents one approach to examine task-oriented conduct, distinct from ordinary conversation, in an attempt to locate and ground the identification of relevant roles. A range of conversational practices for (1) managing agreement/disagreement, (2) managing epistemic authority and subordination, and (3) designing questions and responses to them can be very precisely linked to social roles and culturally variable identities in specific interactions.

Previous work on speaker role detection has mainly concentrated on identifying roles in formal multi-party meetings, broadcast news and broadcast conversations (such as talk shows). Note that most of these conversations have clearly defined roles, such as anchor vs. guest, or professor vs. student. In other words, the roles of the speakers do not change throughout the interaction. This is in contrast to discourse specific roles, where each person may be classified with multiple roles in a conversation. Some typical examples include attacker vs. supporter in a discussion or presenter and listener in a meeting. While there is some work on detecting these discourse specific roles, (e.g., (Banerjee and Rudnicky 2004)), in this section, our focus is on institutional roles.

The earliest work is in this area is only a decade old. Barzilay et al. (2000) focused on identifying three roles in radio broadcast news: the anchor, a journalist, and a guest. Only textual features like word n-grams, explicit speaker introductions, duration features, and features from surrounding context were used. They observed significant gains using the content, but the assumption is that the system knows who is talking when. Later, Liu (2006) studied these three roles in Mandarin broadcast news shows, using a maximum entropy classifier with similar features. She also used contextual speaker role information as a feature.

A study worth discussion about detecting discourse specific roles is by Banerjee and Rudnicky (2004). They have defined three main meeting states (discourse): presentation, discussion, briefing, and other (for capturing smalltalk, etc.). For each meeting state they defined the social roles. For example presenter and listener for the presentation state. They performed experiments using a small in-house meeting data collection. The features they

used include turn-taking patterns, the number of persons speaking at around the same time, and overlap patterns, and no lexical information is exploited.

Vinciarelli (2007) used interaction patterns and information from social network analysis to detect six roles in Swiss radio broadcast news. In this work, information from interaction patterns was captured by centrality measures and relative interactions between speakers, which is computed using duration related features, ignoring the content. Later, Garg et al. (2008) combined lexical features in the form of word n-grams, interaction patterns, and centrality features for identifying speaker roles in multi-party meetings. One interesting observation is that lexical features performed significantly better than only speech-based social network analysis features and they experimentally showed that the combination obtained better results than either type of feature alone using the AMI meeting corpus.

More recently, Yaman et al. (2010) presented a dynamic Bayesian network-based approach for detecting roles in broadcast shows. The main advantage of this work is to capture the sequential patterns of roles between host, guest, journalist, and the audience participants. The features used in the states include the duration of linguistic phenomena in a given turn, and the ratio of the linguistic phenomena to the entire turn, where the linguistic phenomena includes person addresses and mentions, disfluencies, and prefaces (e.g., "Well ...")

Hutchinson et al. (2010) presented an approach for unsupervised speaker role labeling in broadcast talk shows. They used both turn taking features and lexical features to cluster the speaker roles. They aim to find signature phrases (such as "welcome back") which are uttered at all shows but by a very few number of people, and conversational phrases (such as "but" or "you know") which are frequent in broadcast conversations but infrequent in broadcast news. They performed experiments using both English and Mandarin data sets. They found out that spectral clustering beat k-means or GMM and turn-taking features outperform lexical features in English and vice versa in Mandarin.

1.9 Modeling Dominance

An area on which automatic social relation detection research has focused is detecting dominant speakers. These are the people dominating the meeting by controlling the flow of the discussion and conversation, by intervening, cutting off others, and raising new points, and furthermore obeyed by the others. In a typical broadcast talk show, the host may be considered as the dominant person as he has the role of coordinating the conversation. Note that, this is different than the person who has the most authority, which often comes with expertise in an area. Following the broadcast talk show example, an expert scientist guest on some technical area is the authority in the conversation but not necessarily dominating it. Similarly while it is known that dominant speakers behave more actively and talk and move more, in certain interactions this may not be the case. Furthermore, the dominant person in a conversation or a multi-party meeting may change throughout the duration of the interaction and this may well align with the topics discussed.

While, modeling dominance has been extensively studied in social sciences literature, recently a few practical studies have been presented towards automatically tagging the most dominant person in meetings using audio (Rienks and Heylen 2006) and visual cues (Jayagopi et al. 2009), but these studies ignored lexical content. According to Rienks and Heylen (2006), dominant speakers try to assert authority by manipulating the group or certain individuals in the group. This research categorized speakers' dominance level as high,

medium, and low using nonverbal features, such as speaking time, and verbal cues, such as number of words spoken.

Jayagopi et al. (2009) presented a multi-modal processing system using audio (speaking length, energy) and video (motion activity) features, and simply classified speakers with the highest feature values as dominant. They also checked the spoken and visual interruptions such as patterns of who is interrupting whom using speech and video features. They built statistical models using SVMs for the AMI meeting corpus. Using the baseline of tagging the most talkative person as dominant achieved an accuracy of 82%. Using the remaining features this increased to 91%, mostly with the help of speech energy and turn taking patterns.

Future work on dominance modeling involves exploiting content in addition to audio and visual cues, such as checking topic patterns and even some simple key phrases as proven to be useful in the speaker role detection research presented above.

1.10 Argument Diagramming

Argument diagramming aims to display a visual representation of the flow and structure of reasoning in conversations, especially in discussions and arguments (Rienks et al. 2005). The utterances and their relationships are tagged with predefined classes representing the characteristics of discussion and argumentation. For example, one utterance may open a new issue and another utterance may elaborate on it in response. Typically in the argument diagrams, utterances are represented via typed nodes, and relationships between pairs of utterances via typed edges connecting two nodes, forming a tree structure for the topics discussed. Figure 1.10 shows the transcription of an excerpt from the AMI corpus with argument diagram annotations, where the participants are discussing the design of a remote control device. The rectangles mark the nodes, and the arcs mark the relations of the argument diagram.

Argument diagrams extracted from meetings can be useful for meeting participants, to help them in following discussions and catch up with arguments, if the maps can be extracted during the meeting.

There also is a wide body of work on annotating and finding discourse structure, mainly focused on text and dialogs. For example, the Penn Discourse Treebank (PDTB) (Miltsakaki et al. 2004) includes manual annotations of explicit (such as as a result) and implicit discourse connectives, as well as their host and anaphoric arguments in the Wall Street Journal articles. The host argument is the argument in which the discourse connective occurs, while the anaphoric argument is the argument related to the host argument by the discourse connective. While automatically finding anaphoric arguments of discourse connectives is similar to argument diagramming, the PDTB annotation mainly includes annotation of discourse connectives, rather than abstract relations as in argument diagrams. Baldridge and Lascarides (2005) aimed at annotating dialogs with rhetorical relations that conform to Segmented Discourse Representation Theory (Lascarides and Asher 2007), using probabilistic context free grammars to extract the discourse structure of these dialogs. Carlson et al. (2003) presented another study on forming a discourse-tagged corpus, and it also describes measures for quality assurance and consistency, found by mapping hierarchical structures into sets of units and measuring annotator agreement.

There is a wide body of work that focuses on visualization of argument maps, as entered by the conversation participants (Fujita et al. 1998, among others). Argument diagrams

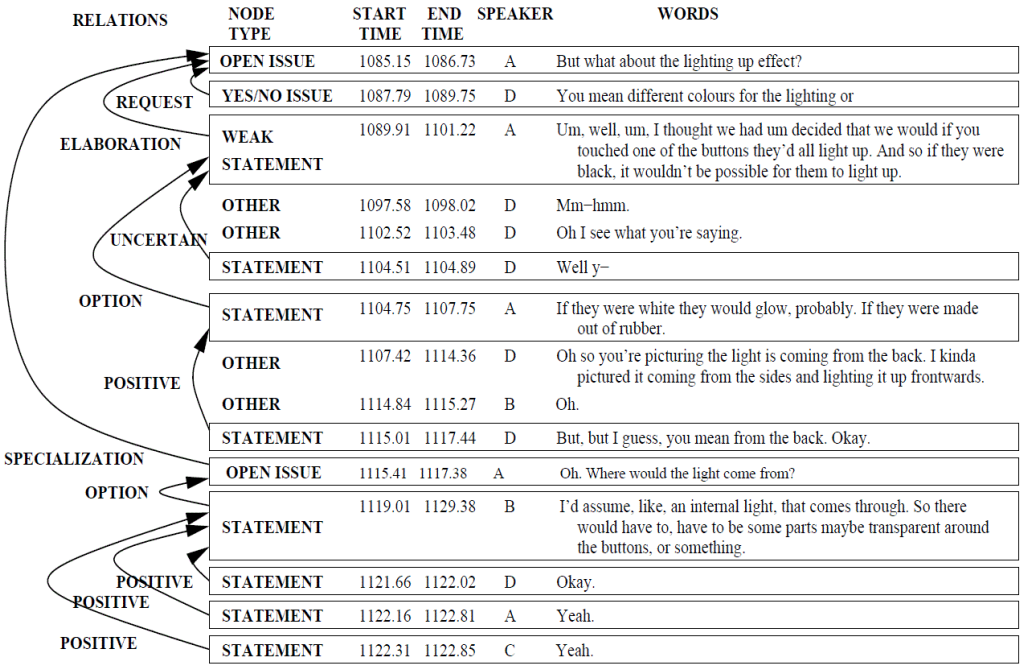

Figure 1.10 Example meeting transcript with argument nodes and relation annotations. The start and end time (in seconds), speaker identity (A,B,C, or D), as well as the words of every speaker turn is shown.

can also help users in browsing past meetings, tracking progress across several meetings and can be useful in meeting summarization. Rienks and Verbree (2006) have performed experiments with human subjects, and their results indicated that argumentation information from meetings can be useful in question answering. Argument diagrams can also help the related tasks of action item extraction and decision detection in meetings (see above). Note that argument diagramming is different than decision detection in several ways, the most important one is that not all discussions are required to include a decision.

For the multiparty meetings domain, two studies proposed argumentative models of meeting discussion. Combining their experience from two meeting processing projects, DARPA CALO and Swiss National Research project IM2, Pallotta et al. (2005) discussed four perspectives (persuasion, decision making, episodes, and conversations), and a theoretical model for each perspective.

Similarly, Rienks et al. (2005) proposed the Twente Argumentation Schema (TAS), and annotated the AMI meeting corpus according to TAS. In this representation, there are six node types, and nine relation types. The relations apply to specific node type pairs. Below we cover TAS in more detail.

TAS was created at University of Twente, where argument diagrams for parts of meeting transcripts that contain discussions around a specific topic, were also formed. In TAS, argument diagrams are tree-structured; the nodes of the tree contain speech act units (usually

TYPE	EXAMPLE
STA	And you keep losing them.
WST	We should probably just use conventional batteries.
OIS	What's the functionality of that?
AIS	So, double or triple?
YIS	Do we need an LCD display?
OTHER	Mm-hmm.

Figure 1.11 Examples of utterances that belong to statement (STA), weak statement (WST), open issue (OIS), A/B issue (AIS), Yes/No issue (YIS), and OTHER node types.

parts of or complete speaker turns) and the edges show the relations between the nodes, the edges emanate from parents and end at children nodes, where the children nodes follow parent nodes in time. At a high level, there are two types of nodes: *issues and statements*. The *issue* nodes mainly open up an issue and request a response and are further categorized into three depending on the form of the response they expect: *open issue* (OIS), *A/B issue* (AIS) and Yes/No issue (YIS). The *open issues* are utterances that allow for various possible responses, that are not included in the utterances themselves. In contrast, *A/B issues* are utterances that request possible responses that are specified in the utterance. The *Yes/No issues* directly request the other participants' opinion as a "Yes" or "No". The *statements* are utterances that convey the position of the speaker on a subject/topic. To be able to represent the statements for which the speaker is not highly certain about what they say, the *statements* are split into two: *statements* (STA) and *weak statements* (WST). The *weak statements* represent the cases where the speaker is not very confident. The rest of the utterances that are not involved in reasoning or backchannelling utterances are represented with an additional (OTHER) category. Table 1.11 shows example utterances for each node type.

The relations between a pair of utterances are categorized into nine types: *Elaboration*, *Specialization*, *Request*, *Positive*, *Negative*, *Uncertain*, *Option*, *Option Exclusion*, and *Subject To*. As its name implies, *Elaboration* relation applies to the pair of utterances (both which can be statements or issues), where the child node utterance elaborates on the parent node utterance. Similarly, the *Specialization* relation applies to pairs (statements and statements or issues and issues), where the child node is a specialization of the parent node. The *Request* relation relates two utterances (statements to issues), where the child utterance asks for more information about the parent. The *Positive* and *Negative* relations apply to utterances, where the child utterance supports or refutes the parent utterance, respectively. The *Uncertain* relation applies to pairs, where it is not clear if the child supports or refutes the parent node. The *Option* relation relates pairs of utterances (statements to issues or other statements), where the where the child is a possible answer, option or solution to the parent utterance. The *Option Exclusion* relates pairs (statements or issues to issues), where the child node eliminates one or more of the possible answers, options or solutions to the parent utterance. The *Subject To* relation applies to pairs (statements and Yes/No or A/B issues or statements), where the child provides criteria that need to be fulfilled before the parent node can be

supported or denied. More information about the relation types and example utterance pairs and annotated tree structures can be found in (Rienks et al. 2005).

Following the TAS schema, Hakkani-Tür (2009) tackled the problem of assigning node types to user utterances, and studied the use of several lexical and prosodic features for this task. More specifically, she has employed a cascaded approach relying on two classifiers using lexical and prosodic features for tagging the argumentation types of the utterances. Prosodic information is shown to be very helpful in distinguishing the backchannels and questions raising issues as expected. One important thing to note about relations is, they usually relate pairs of utterances of specific node types. Therefore, the detection of node types before determining the relations is intuitively the processing sequence for extracting argument diagrams from conversations, while joint modeling techniques should also be investigated in the future.

Murray et al. (2006) investigated the use of prosodic features to detect rhetorical relations, that aim to describe conversations in terms of coherence. Rienks and Verbree (2006) used decision trees with features extracted from manual annotations, such as the presence of a question mark, utterance length, label of the preceding segment, and automatically computed features such as part of speech tags to investigate the learnability of argument diagram node types. While their work resulted in promising improvements over a fairly simple baseline, most of the features used in that work are extracted from manual annotations. Also, the automatic annotation of relations was not considered.

1.11 Discussion and Conclusions

As the enabling speech and language processing technologies are getting more mature and robust, we are in that phase where exploring automatic methods for processing human/human conversational understanding tasks are now feasible. Progress in these tasks, from low-level transcription to higher-level shallow understanding functions, such as action item extraction and summarization, has a potentially enormous impact on human productivity in many professional settings. However, these tasks are still very shallow and focused on targeted understanding of certain phenomena. Most higher level semantic understanding tasks are only vaguely defined and the annotator agreements are still very low. One potential solution is studying limited domain and maybe goal-oriented conversations instead of unstructured chit-chat for better interannotator agreement and hence potentially deeper understanding.

Promising future work includes integration of these tasks and features from multiple modalities, such as from video, or digital pen and paper. Furthermore, personalization of these tasks and exploiting meta information such as project related documentation or emails may bring these technologies to indivdual users. Another interesting research direction would be processing aggregate of conversations and meetings, tracking the topics, participants, and action items, similar to emails.

References

Allen JF, Schubert LK, Ferguson G, Heeman P, Hwang T, Kato T, Light M, Martin NG, Miller BW, Poesio DR and Traum DR 1995 The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical AI (JETAI)* **2**(2), 119–129.

AMI n.d. Augmented multi-party interaction. http://www.amiproject.org.

Anderson A, Bader M, Bard E, Boyle E, Doherty GM, Garrod S, Isard S, Kowtko J, McAllister J, Miller J, Sotillo C, Thompson H and Weinert R 1991 The HCRC maptask corpus. *Language and Speech* **34**, 351–366.

Ang J, Liu Y and Shriberg E 2005 Automatic dialog act segmentation and classification in multiparty meetings *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, PA.

Austin JL 1962 *How to do things with words*. Harvard University Press.

Baldridge J and Lascarides A 2005 Probabilistic head-driven parsing for discourse structure *Proceedings of the CONLL*.

Banerjee S and Rudnicky A 2004 Using simple speech based features to detect the state of a meeting and the roles of the meeting participants *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Jeju-Island, Korea.

Banerjee S, Rosé C and Rudnicky A 2005 The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing *Proceedings of the 10th International Conference on Human-Computer Interaction (CHI)*.

Barzilay R, Collins M, Hirschberg J and Whittaker S 2000 The rules behind roles: Identifying speaker role in radio broadcasts *Proceedings of the Conference of the American Association for Artificial Intelligence (AAAI)*.

Bennett PN and Carbonell J 2005 Detecting action-items in e-mail *Proceedings of the ACM Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Salvador, Brazil.

Bennett PN and Carbonell JG 2007 Combining probability-based rankers for action-item detection *Proceedings of the HLT/NAACL*, pp. 324–331. Association for Computational Linguistics, Rochester, NY.

Bikel DM, Schwartz R and Weischedel RM 1999 An algorithm that learns what's in a name. *Machine Learning Journal Special Issue on Natural Language Learning* **34**(1-3), 211–231.

Bilmes JA and Kirchhoff K 2003 Factored language models and generalized parallel backoff *Proceedings of the Human Language Technology Conference (HLT)-Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Edmonton, Canada.

Burger S, MacLaren V and Yu H 2002 The ISL Meeting Corpus : The impact of meeting type on speech style *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, Colorado.

Carlson L, Marcu D and Okurowski ME 2003 Current directions in discourse and dialogue *Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory*. Kluwer Academic Publishers.

Cetin O and Shriberg E 2006 Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 293–296.

CHI n.d. Computers in the human interaction loop. http://chil.server.de.

Church KW 1988 A stochastic parts program and noun phrase parser for unrestricted text *Proceedings of the Conference on Applied Natural Language Processing (ANLP)*, pp. 136–143, Austin, Texas.

Clark A and Popescu-Belis A 2004 Multi-level dialogue act tags *Proceedings of the SigDial Workshop*, Boston, MA.

Clark HH and Schaefer EF 1989 Contributing to discourse *Cognitive Science*, vol. 13, pp. 259–294.

Core M and Allen J 1997 Coding dialogs with the DAMSL annotation scheme *Proceedings of the Working Notes of the Conference of the American Association for Artificial Intelligence (AAAI) Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA.

Corston-Oliver S, Ringger E, Gamon M and Campbell R 2004 Task-focused summarization of email *Proceedings of the ACL Workshop Text Summarization Branches Out*.

Cuendet S, Hakkani-Tür D and Tur G 2006 Model adaptation for sentence segmentation from speech *Proceedings of the IEEE Spoken Language Technologies (SLT) Workshop*, Aruba.

Dhillon R, Bhagat S, Carvey H and Shriberg E 2004 Meeting recorder project: Dialog act labeling guide. Technical Report TR-04-002, International Computer Science Institute, Berkeley, CA.

Drew P and Heritage J 1992 *Talk at Work*. Cambridge University Press.

Favre B, Grishman R, Hillard D, Ji H and Hakkani-Tür D 2008a Punctuating speech for information extraction *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV.

Favre B, Hakkani-Tür D, Petrov S and Klein D 2008b Efficient sentence segmentation using syntactic features *Proceedings of the IEEE Spoken Language Technologies (SLT) Workshop*, Goa, India.

Fernández R, Frampton M, Ehlen P, Purver M and Peters S 2008 Modelling and detecting decisions in multi-party dialogue *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pp. 156–163. Association for Computational Linguistics, Columbus, OH.

Frampton M, Fernández R, Ehlen P, Christoudias M, Darrell T and Peters S 2009a Who is "you "? combining linguistic and gaze features to resolve second-person references in dialogue *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Frampton M, Huang J, Bui TH and Peters S 2009b Real-time decision detection in multi-party dialogue *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

Fujita K, Nishimoto K, Sumi Y, Kunifuji S and Mase K 1998 Meeting support by visualizing discussion structure and semantics *Proceedings of the Second International Conference on Knowledge-Based Intelligent Electronic Systems*, Adelaide, Australia.

Galley M, McKeown K, Hirschberg J and Shriberg E 2004 Identifying agreement and disagreement in conversational speech: Use of B ayesian networks to model pragmatic dependencies *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Garg N, Favre S, Salamin H, Hakkani-Tür D and Vinciarelli A 2008 Role recognition for meeting participants: an approach based on lexical information and social network analysis *Proceedings of the ACM Multimedia Conference*.

Godfrey JJ, Holliman EC and McDaniel J 1992 Switchboard: Telephone speech corpus for research and development *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 517–520, San Francisco, USA.

Gupta S, Niekrasz J, Purver M and Jurafsky D 2007 Resolving "you" in multi-party dialog *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Guz U, Favre B, Tur G and Hakkani-Tür D 2009 Generative and discriminative methods using morphological information for sentence segmentation of Turkish . *IEEE Transactions on Audio, Speech, and Language Processing* **17**(5), 895–903.

Guz U, Tur G, Hakkani-Tür D and Cuendet S 2010 Cascaded model adaptation for dialog act segmentation and tagging. *Computer Speech and Language* **18**(2), 289–306.

Hakkani-Tür D 2009 Towards automatic argument diagramming of multiparty meetings *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan.

Hillard D, Ostendorf M and Shriberg E 2003 Detection of agreement vs. disagreement in meetings: Training with unlabeled data *Companion Volume of the Proceedings of the HLT-NAACL -Short Papers*, Edmonton, Alberta.

Hillard D, Ostendorf M, Stolcke A, Liu Y and Shriberg E 2004 Improving automatic sentence boundary detection with confusion networks *Proceedings of the Human Language Technology Conference (HLT)-NAACL*, Boston, MA.

Hsueh PY and Moore J 2007 What decisions have you made?: Automatic decision detection in meeting conversations *Proceedings of NAACL/HLT*, Rochester, New York.

Hutchinson B, Zhang B and Ostendorf M 2010 Unsupervised broadcast conversation speaker role labeling *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Janin A, Ang J, Bhagat S, Dhillon R, Edwards J, Macias-Guarasa J, Morgan N, Peskin B, Shriberg E, Stolcke A, Wooters C and Wrede B 2004 The ICSI meeting project: Resources and research *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal.

Jayagopi DB, Hung H, Yeo C and Gatica-perez D 2009 Modeling dominance in group conversations using non-verbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing* **17**(3), 501–513.

Jovanovic N, op den Akker R and Nijholt A 2006 Addressee identification in face-to-face meetings *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 169–176, Trento, Italy.

Jurafsky D and Martin JH 2008 *Speech and Language Processing* second edition edn. Prentice Hall, NJ.

Jurafsky D, Shriberg E and Biasca D 1997 Switchboard SWBD-DAMSL labeling project coder's manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science.

Kolar J, Liu Y and Shriberg E 2007 Speaker adaptation of language models for automatic dialog act segmentation of meetings *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, Antwerp, Belgium.

Kolar J, Shriberg E and Liu Y 2006 Using prosody for automatic sentence segmentation of multi-party meetings *Proceedings of the International Conference on Text, Speech, and Dialogue (TSD)* , Czech Republic.

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

Lascarides A and Asher N 2007 Segmented discourse representation theory: Dynamic semantics with discourse structure In *Computing Meaning* (ed. Bunt H and Muskens R). Kluwer Academic Publishers.

Lisowska A 2003 Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical Report IM2.MDM-11 , ISSCO , University of Geneva.

Liu Y 2006 Initial study in automatic identification of speaker role in broadcast news speech *Proceedings of the Human Language Technology Conference (HLT) / Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New York City, USA.

Liu Y and Xie S 2008 Impact of automatic sentence segmentation on meeting summarization *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV.

Liu Y, Stolcke A, Shriberg E and Harper M 2005 Using conditional random fields for sentence boundary detection in speech *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI.

Margolis A, Livescu K and Ostendorf M 2010 Domain adaptation with unlabeled data for dialog act tagging *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Mast M, Kompe R, Harbeck S, Kiessling A, Niemann H, Nöth E, Schukat-Talamazzini EG and Warnke V 1996 Dialog act classification with the help of prosody *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia.

Matusov E, Hillard D, Magimai-Doss M, Hakkani-Tür D, Ostendorf M and Ney H 2007 Improving speech translation with automatic boundary prediction *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, Antwerp, Belgium.

Miltsakaki E, Prasad R, Joshi A, and Webber B 2004 The penn discourse treebank *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Morgan W, Chang PC, Gupta S and Brenier JM 2006 Automatically detecting action items in audio meeting recordings *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pp. 96–103. Association for Computational Linguistics, Sydney, Australia.

Müller C 2006 Automatic detection of nonreferential *It* in spoken multi-party dialog *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 49–56, Trento, Italy.

Müller C 2007 Resolving it, this, and that in unrestricted multi-party dialog *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 816–823.

Murray G, Renals S and Taboada M 2006 Prosodic correlates of rhetorical relations *Proceedings of Human Language Technology Conference (HLT)-NAACL Workshop on Analyzing Conversations in Text and Speech (ACTS)*, New York City, NY, USA.

Pallotta V, Niekrasz J and Purver M 2005 Collaborative and argumentative models of meeting discussions *Proceedings of the 5th Workshop on Computational Models of Natural Argument (CMNA)*, Edinburgh, Scotland.

Pang B and Lee L 2008 *Opinion mining and sentiment analysis*. Now publishers.

Polifroni J, Seneff S, Branavan SRK, Wang C and Barzilay R 2010 Good grief, i can speak it! preliminary experiments in audio restaurant reviews *Proceedings of the IEEE Spoken Language Technologies (SLT) Workshop*, Berkeley, CA.

Popescu-Belis A 2005 Dialogue acts: One or more dimensions. ISSCO Working Paper n. 62, University of Geneva.

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

Purver M, Dowding J, Niekrasz J, Ehlen P, Noorbaloochi S and Peters S 2007 Detecting and summarizing action items in multi-party dialogue *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Purver M, Fernández R, Frampton M and Peters S 2009 Cascaded lexicalised classifiers for second-person reference resolution *Proceedings of the SIGDIAL Meeting on Discourse and Dialogue,*, London, UK.

Rangarajan V, Bangalore S and Narayanan S 2007 Exploiting prosodic features for dialog act tagging in a discriminative modeling framework *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, Antwerp, Belgium.

Rienks R and Heylen D 2006 Automatic dominance detection in meetings using easily obtainable features In *MLMI, Revised Selected Papers* (ed. Renals S and Bengio S) vol. 3869 of *Lecture Notes in Computer Science* Springer pp. 76–86.

Rienks R and Verbree D 2006 About the usefulness and learnability of argument-diagrams from real discussions *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Washington D.C., USA.

Rienks R, Heylen D and van der Weijden E 2005 Argument diagramming of meeting conversations *Multimodal Multiparty Meeting Processing, Workshop at the International Conference on Multimodal Interaction (ICMI)*, Trento, Italy.

Roark B, Liu Y, Harper M, Stewart R, Lease M, Snover M, Shafran I, Dorr B, Hale J, Krasnyanskaya A and Yung L 2006 Reranking for sentence boundary detection in conversational speech *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France.

Schuller B, Wöllmer M, Eyben F and Rigoll G 2010 Retrieval of paralinguistic information in broadcasts In *Multimedia Information Extraction* (ed. Maybury M). MIT Press, Cambridge, MA.

Shriberg E, Bates R, Stolcke A, Taylor P, Jurafsky D, Ries K, Coccaro N, Martin R, Meteer M and Ess-Dykema CV n.d. Can prosody aid the automatic classification of dialog acts in conversational speech?

Shriberg E, Dhillon R, Bhagat S, Ang J and Carvey H 2004a The ICSI Meeting Recorder Dialog Act (MRDA) corpus *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at Human Language Technology Conference (HLT)-NAACL 2004*, pp. 97–100.

Shriberg E, Dhillon R, Bhagat S, Ang J and Carvey H 2004b The ICSI Meeting Recorder Dialog Act (MRDA) Corpus *Proceedings of the SigDial Workshop*, Boston, MA.

Shriberg E, Stolcke A, Hakkani-Tür D and Tur G 2000 Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* **32**(1-2), 127–154.

Somasundaran S, Ruppenhofer J and Wiebe J 2007 Detecting arguing and sentiment in meetings *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium.

Somasundaran S, Wiebe J, Hoffmann P and Litman D 2006 Manual annotation of opinion categories in meetings *Proceedings of the ACL/COLING Workshop: Frontiers in Linguistically Annotated Corpora*, Sydney, Australia.

Stent A 2000 The monroe corpus. TR728 and TN99-2, University of Rochester.

Stiefelhagen R, Yang J and Waibel A 2002 Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks* **13**(3), 928–938.

Stolcke A and Shriberg E 1996 Statistical language modeling for speech disfluencies *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA.

Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, van Ess-Dykema C and Meteer M 2000 Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics* **26**(3), 339–373.

Susanne J, Klein A, Maier E, Maleck I, Mast M and Quantz J 1995 Dialogue acts in VERBMOBIL . Report 65, University of Hamburg, DFKI GmbH, University of Erlangen, and TU Berlin.

DARPA Cognitive Agent that Learns and Organizes (CALO) Project

DARPA Cognitive Agent that Learns and Organizes (CALO) Project n.d. http://www.ai.sri.com/project/CALO.

Traum DR and Hinkelman EA 1992 Conversation acts in task-oriented spoken dialogue. *Computational Intelligence* **8**(3), 575–599.

Tur G, Guz U and Hakkani-Tür D 2006 Model adaptation for dialog act tagging *Proceedings of the IEEE Spoken Language Technologies (SLT) Workshop*.

Tur G, Stolcke A, Voss L, Peters S, Hakkani-Tür D, Dowding J, Favre B, Fernandez R, Frampton M, Frandsen M, Frederickson C, Graciarena M, Kintzing D, Leveque K, Mason S, Niekrasz J, Purver M, Riedhammer K, Shriberg E, Tien J, Vergyri D and Yang F 2010 The CALO meeting assistant system. *IEEE Transactions on Audio, Speech, and Audio Processing* **18**(6), 1601–1611.

Venkataraman A, Liu Y, Shriberg E and Stolcke A 2005 Does active learning help automatic dialog act tagging in meeting data? *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, Lisbon, Portugal.

Venkataraman A, Stolcke A and Shriberg EE 2002 Automatic dialog act tagging with minimal supervision *Proceedings of the Australian International Conference on Speech Science and Technology*, Melbourne, Australia.

Verbree A, Rienks R and Heylen D 2006 First steps towards the automatic construction of argument-diagrams from real discussions *Proceedings of the 1st International Conference on Computational Models of Argument, September 11 2006, Frontiers in Artificial Intelligence and Applications*, vol. 144, pp. 183–194. IOS press.

Vinciarelli A 2007 Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia Processing*.

Warnke V, Kompe R, Niemann H and Nöth E 1997 Integrated dialog act segmentation and classification using prosodic features and language models *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece.

Wilson T 2008 Annotating subjective content in meetings *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Wrede B and Shriberg E 2003 The relationship between dialogue acts and hot spots in meetings *Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*.

Wrede B, Bhagat S, Dhillon R and Shriberg E 2005 Meeting recorder project: Hot spot labeling guide. Technical Report TR-05-004, International Computer Science Institute, Berkeley, CA.

Yaman S, Hakkani-Tür D and Tur G 2010 Social role discovery from spoken language using dynamic bayesian networks *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, Makuhari, Japan.

Yang F, Tur G and Shriberg E 2008 Exploiting dialog act tagging and prosodic information for action item identification *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV.

Zimmerman M, Hakkani-Tür D, Fung J, Mirghafori N, Gottlieb L, Shriberg E and Liu Y 2006 The ICSI+ multilingual sentence segmentation system *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburg, PA.

Zimmermann M, Liu Y, Shriberg E and Stolcke A 2005 Toward joint segmentation and classification of dialog acts in multiparty meetings *Proceedings of the Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, Edinburgh, U.K.

Zue V, Seneff S and Glass J 1990 Speech database development at MIT : Timit and beyond. *Speech Communication* **9**(4), 351–356.