Structured Probabilistic Modelling for Dialogue Management

Doctoral Dissertation by

Pierre Lison

Department of Informatics
Faculty of Mathematics and Natural Sciences
University of Oslo

Abstract

TODO

Acknowledgements

TODO

don't forget to mention the HPC infrastructure and the YRRSDS gang

Contents

Chapter 1

Introduction

Spoken language is one of the most powerful system of communication at our disposal. A large part of our waking hours is spent in social interactions mediated through natural language. The pivotal role of spoken language in our daily lives is largely due to its remarkable proficiency at conveying elaborate thoughts in a robust, flexible and efficient manner.

Is it possible to exploit this simple observation to develop more human-friendly technologies? Most of our everyday activities are now relying on "smart" electronic devices of various kinds, from mobile phones to personal computers and in-car navigation systems. As these technologies gain in autonomy and sophistication, it becomes increasingly important to design user interfaces that can combine expressiveness, adaptivity and ease of use. In this context, it seems judicious to endow these devices with a capacity to understand, even in a limited manner, the communication medium that is most natural to us, namely spoken language.

The ongoing research on *spoken dialogue systems* (SDS) is precisely trying to achieve this objective. A spoken dialogue system is a computer agent that is able to converse with humans through everyday spoken language in order to perform its task(s). Such systems are expected to play an ever-increasing role in our daily interactions with technology. They have a wide range of applications, ranging from phone-based systems for information access and service delivery to voice-enabled software for hand-held devices, navigation assistants, interactive tutoring systems, and (in a not-too-distant future) service robots assisting us in our everyday environments.

Figure 1.1 illustrates an example of interaction between a human user and a spoken dialogue system. When the user starts talking, the system extracts the corresponding speech signal through a microphone. The speech signal is then processed to analyse its content. Once this operation is completed, the system must then decide how to react. In our case, the system decides to greet back the user and selects the words to express it ("*good morning, sir*"). The final step is then to synthesise these words through an artificial voice, which closes the loop[1].

1.1 Motivation

Although the deployment of spoken dialogue systems is attractive for many reasons, their practical development can be a demanding enterprise. Speech is indeed much more complex than other

---

[1] Needless to say, the schema hides a great deal of internal complexity. In particular, it omits the existence of non-verbal inputs and outputs (e.g. additional modalities, external actions) which are present in most applications. The next chapter will describe in more details the software architectures used to design spoken dialogue systems.
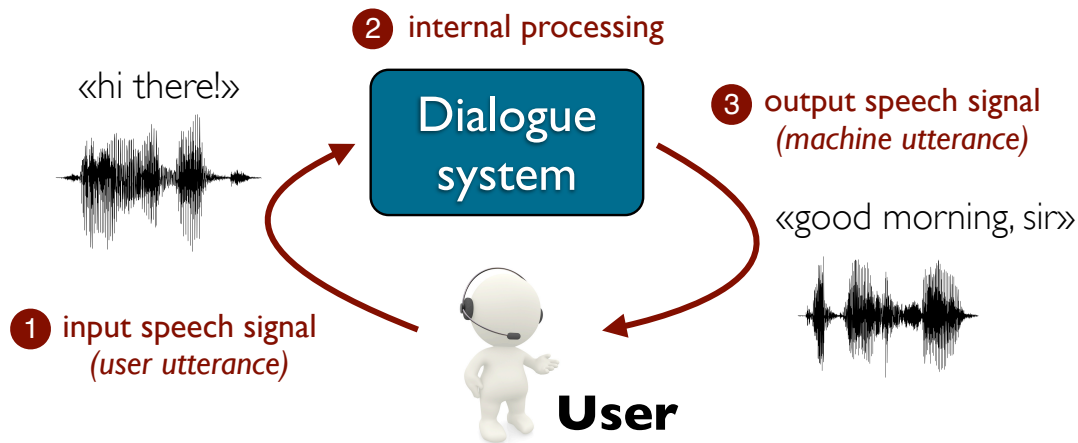
Figure 1.1: Schematic view of a spoken dialogue system

modalities for user interaction such as keyboards or touch screens.

The present thesis concentrates on the problem of *dialogue management*. Dialogue management is a central function in spoken dialogue systems. It serves a double role. Its first task is to maintain a representation of the current dialogue state. This representation might include any information that is relevant for the system, and often include features related to the dialogue history, the external context, and the current tasks to perform. This dialogue state is regularly updated with new information, which comes either in the form of new user utterances or perceived changes in the context. The second task of dialogue management is to make decisions. Based on the current state of the interaction, dialogue management must decide which actions to undertake. These actions are often communicative in nature (e.g. uttering a sentence), but can also pertain to physical actions to execute (e.g. grasping an object).

Dialogue management is therefore responsible for controlling the flow of the interaction, by (1) deciding how to interpret the user inputs in their context and (2) selecting which actions to perform next. In the example from Figure 1.1, this step corresponds to the decision of responding to the user utterance "*hi there!*" with another greeting action, "*good morning, sir*".

Along with speech recognition, dialogue management has proven be be one of the most difficult processing component in spoken dialogue systems. To understand why this is the case, it is useful to have a closer look at two defining features of verbal interactions:

1. Verbal interactions are highly *structured*. Their analysis reveals the presence of multiple relations straddling the linguistic and extra-linguistic levels of interpretation. Spontaneous conversations are notably rife with partial utterances and referring expressions that can only be uncovered based on their context. Furthermore, verbal interactions are at their core a collaborative activity where the contributions of the different speakers are tightly coupled (Clark, 1996; Garrod and Pickering, 2004). A dialogue move is therefore only intelligible within the larger pragmatic context that gave rise to it.

2. Verbal interactions are also crippled with *uncertainties*. In order to make sense of a given utterance, a conversational agent must face numerous sources of uncertainty, including error-prone speech recognition, lexical, syntactic and referential ambiguities, partially observable environments, and unpredictable interaction dynamics.

The combination of these two properties forms an explosive mix. In order to make sense of the interaction and act appropriately, the dialogue system must be able to perform complex reasoning in order to interpret the user intentions and plan the best course of action. And it must do so under high levels of noise and uncertainty, where many pieces of information can contain errors or be missing, ambiguous, or fragmentary. This problem is known in Artificial Intelligence as *sequential decision-making under uncertainty* (Kaelbling et al., 1998; Russell and Norvig, 2010), and it remains to this day a difficult computational task, especially for complex domains such as dialogue.

Research on dialogue management can be divided into two main lines of investigation that reflect their focus on either of the two linguistic properties we just mentioned.

On the one hand, structural complexity is often dealt with conceptual tools borrowed from formal logic. These approaches provide principled methods for the interpretation and generation of dialogue moves through logical reasoning on the basis of a formal representation of the mental states of the dialogue participants (including their shared knowledge). This representation might incorporate the beliefs, desires and intentions of each agent (Cohen and Perrault, 1979; Allen and Perrault, 1980), social obligations (Traum and Allen, 1994), or open questions raised during the interaction (Larsson, 2002; Ginzburg, 2012). These approaches can provide detailed analyses of various dialogue behaviours, but they generally assume complete observability of the dialogue context and provide only a very limited account (if any) of errors and uncertainties. In addition, they require the knowledge base on which the inference is grounded to be completely specified in advance by domain experts. Their deployment in practical applications is therefore non trivial.

On the other hand, the problem of uncertainty is usually addressed by probabilistic modelling techniques (Roy et al., 2000; Frampton and Lemon, 2009; Young et al., 2010). The state of the dialogue is here represented as a probability distribution over possible worlds. This distribution represents the system's current knowledge of the interaction and is regularly updated as new observations are collected. These probabilistic models provide an explicit account for the various uncertainties that can arise during the interaction. They also enable the dialogue behaviour to be automatically optimised in a data-driven manner instead of relying on hand-crafted mechanisms. Dialogue strategies can therefore be adapted to new environments or users without having to be reprogrammed. However, these models typically depend on large amounts of training data to estimate their parameters – a requirement that is hard to satisfy for most dialogue domains. In addition, the probabilistic models are usually limited to a handful of state variables and are difficult to scale to domains featuring rich conversational contexts.

The work described in this thesis aims at reconciling these two strands of research via a new, hybrid computational framework for dialogue modelling.

1.2   Contributions

The present thesis details an original approach to dialogue management based on *structured probabilistic modelling*. Its objective is to provide probabilistic models of dialogue that are (1) more scalable to rich domains, (2) easier to estimate from small amounts of training data.

There is an extensive body of work in the machine learning and decision-theoretic planning literature which shows how to address this issue by relying on more expressive representations, able to capture relevant aspects of the problem *structure* in a compact manner. By taking advantage of hierarchical or relational abstractions, system developers can leverage their domain knowledge to

yield probabilistic models which are both easier to learn (due to a reduced number of parameters) and more efficient to use (since the structure can be exploited by the inference algorithm).

This thesis demonstrates how to translate these insights into dialogue modelling. We present a new framework for describing probabilistic models of dialogue, based on the concept of *probabilistic rules*. These rules express the distributions in terms of structured mappings associating specific conditions on a set of input variables to probabilistic effects defined on a set of output variables.

The presented modelling framework offers two major benefits. Most importantly, the reliance on more expressive representations allows us to drastically reduce the number of parameters associated with the models. Instead of being encoded through traditional probability tables, the conditional distributions between variables are expressed through high-level rules that capture the dependencies between variables via a small set of parameters (one for each possible effect). As a consequence, these models are much easier to learn and generalise to unseen data.

In addition, the framework enables expert knowledge to be directly incorporated into the probabilistic models. System developers are thus free to exploit powerful abstractions to encode their prior knowledge of the dialogue domain in the form of pragmatic rules, generic background knowledge, or task-specific constraints. While there exists previous work on the integration of expert knowledge using finite-state policies or ad-hoc constraints (Heeman, 2007; Williams, 2008), these approaches essentially use this information source as an external filter to a classical model. By contrast, our approach seeks to incorporate this expert knowledge in the very structure of the statistical model.

At runtime, these probabilistic rules are instantiated on the variables that compose the dialogue state. This instantiation is realised by converting the rules into the nodes of a Bayesian Network (a.k.a. a directed graphical model). The probabilistic rules can therefore be seen as providing high-level templates for the construction of a classical probabilistic model. After grounding the rules in the Bayesian Network, various inference operations can be triggered to e.g. update the model with new observations or search for the action yielding the highest utility. Standard algorithms for probabilistic inference such variable elimination (Zhang and Poole, 1996) and likelihood weighted sampling (Fung and Chang, 1989) are applied for this purpose.

We conducted several experiments to assess the validity of our approach in different learning scenarios:

1. The first experiment, detailed in Section 4.3, focussed on the problem of estimating the utilities of various system actions given a small data set collected from Wizard-of-Oz interactions.[2]. Based on dialogue models encoded with probabilistic rules, the utilities of the different actions were learned through the systematic application of Bayesian inference in a supervised learning setting. We were then able to show that the rule structure enabled the learning algorithm to converge faster and with better generalisation performance than unstructured models. This work was originally presented in (Lison, 2012b).

2. The second experiment, described in Section 5.3, extended the above approach to reinforcement learning. The goal of this study was to estimate the transition model of the domain

---

[2]A Wizard-of-Oz interaction is an experimental procedure borrowed from Human-Computer Interaction (HCI) studies (Dahlbäck et al., 1993). In a Wizard-of-Oz experiment, the subjects are asked to interact with a computer system which has all the appearances of reality, but is actually remotely controlled by an (unseen) human agent operating behind the curtains. Wizard-of-Oz studies are often conducted to provide the system designers with interaction data from real users before the system is fully implemented.

from interactions with a user simulator. We compared the relative learning performance of two modelling approaches: one relying on unstructured distributions, and one based on probabilistic rules. The empirical results demonstrated once more the benefits of capturing the domain structure with probabilistic rules. The results were first published in XXX

3. Finally, the third experiment was designed to evaluate the approach through live interactions with real users. to be completed

An additional contribution of our thesis is a software toolkit that implements all the representations and algorithms presented in this work. The toolkit is dubbed `openDial` and is freely available under an open source licence.[3] It enables system developers to design, evaluate and deploy dialogue systems based on probabilistic rules. The toolkit is fully generic since all all domain-specific knowledge is declaratively specified in the rules for the domain. This design choice effectively simplifies the system architecture to a small set of core algorithms for accessing and updating the dialogue state (Lison, 2012a). The `openDial` toolkit comes with a user interface allowing developers to interactively test their system and visualise how the internal dialogue state is evolving over time. Its implementation is described in Appendix C.

We carried out all the experiments described in this thesis in a *human–robot interaction* (HRI) domain. The selection of this specific domain as a test bed for our framework was motivated by two factors. First of all, HRI domains often embody a rich mix of contextual features extracted from the situated environment and the tasks to complete by the agent. Moreover, HRI domains must frequently significant levels of uncertainty due to imperfect sensors, unreliable motors, and failure-prone speech recognition.

Figure 1.2: Human user interacting with the Nao robot.

The Nao robot from Aldebaran Robotics was used as a platform for all our experiments [4]. An example of interaction with the robot is shown in Figure 1.2. Most of our experiments involved the Nao robot interacting with a human user in a shared visual environment featuring a few basic objects that can be automatically perceived by the robot. The user were instructed to command the robot to execute various tasks such as grasping objects and moving them from one place to the other. The robot was also able to answer questions related to his own perception (e.g. "*do you see the red cylinder?*"). A detailed description of the experimental setup will be provided in the upcoming chapters.
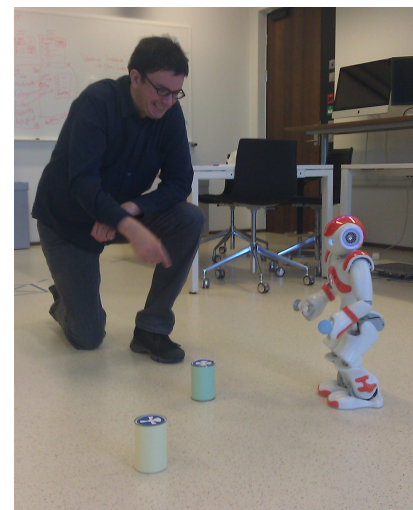
1.3 Outline of the Thesis

We provide here a brief outline of the thesis structure, chapter by chapter.

---

[3]The toolkit can be downloaded at http://opendial.googlecode.com.
[4]cf. http://www.aldebaran-robotics.com.

Chapter 2: Background

This chapter introduces the fundamental concepts and methods used throughout this thesis. We start with an overview of some of the core linguistic properties of dialogue and describe key notions such as turn-taking, dialogue acts and grounding. We then describe the software architectures used to design spoken dialogue systems and the role of each component within them. We also mention a range of important applications for spoken dialogue systems. Finally, we survey the various approaches that have been put forward in the research literature to address the dialogue management problem. In particular, we review both hand-crafted and statistical approaches to the design of dialogue strategies.

Chapter 3: Probabilistic Rules

This chapter lays down the theoretical foundations of our approach. We start by reviewing the core notions of graphical models, since they constitute the formal basis for our framework. We then define what probabilistic rules are and how they are internally structured through conditions and effects. We describe two main types of rules, used to respectively encode probability and utility models. Following this, we explain how the rules are practically instantiated in the Bayesian Network representing the dialogue state. The chapter also addresses some advanced modelling questions, and concludes by discussing related work that also aimed at reducing the dimensionality problem when learning dialogue strategies.

Chapter 4: Learning from Wizard-of-Oz data

This chapter shows how the parameters attached to probabilistic rules can be automatically learned from training data, in a supervised learning fashion. The algorithm to estimate these parameters is grounded in Bayesian inference. To validate our approach, we detail an experiment showing how to learn the utilities of a set of actions from Wizard-of-Oz data collected in a human–robot interaction domain. The experiment illustrates in particular the benefits of applying probabilistic rules.

Chapter 5: Learning from Interactions

This chapter builds upon the previous chapter and extends it to a reinforcement learning context. We show that is is possible to efficiently learn the parameters of dialogue models from observations collected during the interaction itself, without having access to any gold standard annotations. The learning procedure follows a model-based Bayesian reinforcement learning approach. Finally, we report the results of an experiment carried out with a user simulator. The experiment concentrated on the estimation of the transition model in a HRI domain, and evaluated the relative performance of a model structured with probabilistic rules compared to a plain probabilistic model.

Chapter 6: User Evaluation

This chapter presents a user evaluation of our approach in a HRI domain. XXX

Chapter 7: Concluding Remarks

The final chapter concludes this dissertation with a summary of the presented research contributions, followed by an outline of future work.

Chapter 2

Background

We introduce in this chapter the most important concepts and methods employed in the field of spoken dialogue systems, with special emphasis on dialogue management. We start by laying down the linguistic foundations of our work and describe some of the key properties of dialogue. A proper understanding of these aspects is indeed a prerequisite for the design of conversationally competent dialogue systems. After this linguistic overview, we move to a more technical discussion of the software architectures used to implement practical dialogue systems. These architectures typically comprise multiple processing components, from speech recognition to understanding, dialogue management, output generation and speech synthesis. We briefly describe the role of each component and its position in the global processing pipeline. Last but not least, the final section of this background chapter delves into the diverse set of approaches that have been explored in the literature to formalise the dialogue management problem. We first present hand-crafted approaches, starting with finite-state policies and pursuing with more sophisticated methods based on logic- or plan-based reasoning. Finally, we detail the more recently developed statistical approaches to dialogue management that aim to automatically extract optimal policies from data.

2.1 What is spoken dialogue?

We communicate in order to fulfil a wide array of social functions, such as exchanging ideas, recollecting experiences, sustaining relationships, or collaborating with others to accomplish shared goals. These communication skills are developed in early childhood, and our cognitive abilities are in many ways shaped and amplified by this disposition for verbal interaction.

One of the most important property of dialogue is that it is fundamentally a *collaborative activity* (emphasis on both terms). It is, first of all, an *activity*, which means that is is (1) driven by concrete goals to achieve; (2) subject to costs – the communication effort – to minimise, and (3) composed of a temporal sequence of basic actions. Furthermore, if we abstract from so-called "internal dialogues" with oneself, dialogue involves per definition at least two participants. As shown by a wealth of studies in psychology and linguistics (Clark and Schaefer, 1989; Allwood et al., 1992; Clark, 1996; Garrod and Pickering, 2004; Tomasello et al., 2005), human conversations are characterised by a high degree of *collaboration* between interlocutors to coordinate their contributions and ensure mutual understanding.

We describe in the next sections three major aspects of this collaborative activity:

1. The dialogue participants take *turns* in a conversation;

2. These turns are structured into basic communicative units called *dialogue acts*;

3. The participants continuously provide *grounding signals* to each other in order to indicate how they understand (or fail to understand) their contributions.

2.1.1   Turn-taking

Turn-taking is one of the most basic (yet often neglected) aspect of spoken dialogue. The physical constraints of the communication channel impose that participants take turns in order to speak. Turn-taking is essentially a resource allocation problem. In this case, the resource to allocate is the conversational floor, and social conventions dictate how the dialogue participants are to take and release their turns.

The field of *conversation analysis* studies what these conventions are and how they combine to shape conversational behaviours in various languages and cultures. Human conversations are indeed remarkably efficient at turn-taking. Empirical cross-linguistic studies have shown that the average transition time between turns revolves around 250 ms. (Stivers et al., 2009).[1] In addition, most of the utterances do not overlap: Levinson (1983) argues that less than 5 % of the speech stream contains some form of overlap in spontaneous conversations.

A wide variety of cues are used to detect turn boundaries, such as silence, hesitation markers, syntax (complete grammatical unit), intonation (rising or falling pitch) and body language (Duncan, 1972). These cues can occur jointly or in isolation. Upon reaching a turn boundary, a set of social conventions govern who is allowed to take the turn. The current speaker can explicitly select the next person to take the turn, for instance when greeting someone or asking a directed question. This selection can also occur via other mechanisms such as gaze (Sacks et al., 1974). When no such selection is indicated, other participants are allowed to take the turn. Alternatively, the current speaker can continue to hold the floor until the next boundary.

Turn-taking is closely related to the notion of *initiative* in research on human–computer interaction. The vast majority of dialogue systems currently deployed are either system-initiated or user-initiated. In a system-initiated dialogue, the dialogue system has full control on how the interaction is unfolding – i.e. the system is the one asking the questions and waiting for the user responses. A user-initiated dialogue is the exact opposite: in such settings, the user is assumed to lead the interaction and request information from the system. The most complex – but also most natural – interaction style is the mixed-initiative, where both the user and the dialogue system are allowed to take the initiative at any time and provide or solicit information whenever they see fit (Horvitz, 1999).

The turn-taking behaviour of most current-day dialogue systems remains quite primitive. The most common method to detect the end of a user turn is to wait for a silence longer than a manually fixed threshold, typically ranging between ½ and 1.0 second. Some system architectures also include routines for handling barge-ins – that is, user interruptions – (Ström and Seneff, 2000), while others simply ignore them altogether . Turn-taking has now become a focus of research in its own right

---

[1]Interestingly, this duration is shorter than the time required for a human speaker to plan the motor routines associated with the physical act of speaking. This means that the next speaker must start planning his utterance before the current turn is complete, and predict when a potential turn boundary is likely to appear.

in the dialogue system literature (Raux and Eskenazi, 2009; Gravano and Hirschberg, 2011), in an effort to break away from the ping-pong interaction style that characterises most current dialogue interfaces.

2.1.2 Dialogue acts

Each turn is constituted of one or more utterances. As argued by Austin (1962) and Searle (1969), our utterances are nearly always purposeful: they have specific goals and are intended to have a specific effect on the listener. They should therefore best be described as actions rather than abstract statements about the world. The notion of dialogue act embodies this idea.[2] Bunt (1996) defines a dialogue act as a "functional unit of a dialogue used by the speaker to change the context".

In his seminal work on the philosophy of language, Searle (1979) established a taxonomy of speech acts divided in five central categories:

Assertives: Committing the speaker to the truth of a proposition.
  Examples: "*I swear I saw him on the crime scene.*", "*I bought more coffee.*"

Directives: Attempts by the speaker to get the addressee to do something.
  Examples: "*Clean your room!*", "*Could you post this for me?*"

Commissives: Committing the speaker to some future course of action.
  Examples: "*I will deliver this review before Monday.*", "*I promise to work on this.*"

Expressives: Expressing the psychological state of the speaker about a state of affairs.
  Examples: "*I am so happy for you!*", "*Apologies for being late.*"

Declaratives: Bringing about a different state of the world by the utterance.
  Examples: "*You're fired.*", "*We decided to let you pass this exam.*"

Modern taxonomies of dialogue acts are significantly more detailed than the one introduced by Searle. They also provide detailed accounts of various dialogue-level phenomena such as grounding (cf. next section) that were absent from Searle's analysis. The most well-known annotation scheme is DAMSL (Dialogue Act Markup in Several Layers), which was initially put forward by Core and Allen (1997). DAMSL defines a rich, multi-layered annotation scheme for dialogue acts that is both domain- and task- independent. A modified version of this scheme was applied to annotate the Switchboard corpus[3] based on a set of 42 distinct dialogue acts (Jurafsky et al., 1997), including greeting and closing actions, acknowledgements, clarification requests, self-talk, responses, and many more. An interesting aspect of DAMSL is the use of two complementary dimensions in the markup: the *forward-looking functions*, which are the traditional speech acts in Searle's sense (assertions, directives, information requests, etc.) and the *backward-looking functions* that respond

---

[2]Dialogue acts have gone through multiple names over time, owing to the diverse range of research fields that have studied them, from philosophy to descriptive and computational linguistics. As listed in McTear (2004), alternative denominations include speech acts (Searle, 1969), communicative acts (Allwood, 1976), conversation acts (Traum and Hinkelman, 1992), conversational moves (Sinclair and Coulthard, 1975), and dialogue moves (Larsson et al., 1999).

[3]The Switchboard corpus is a corpus of spontaneous telephone conversations collected in the early 1990's. It includes about 2430 conversations averaging 6 minutes in length; totalling over 240 hours of recorded speech with native speakers of American English (Godfrey et al., 1992).

back to a previous dialogue act and can signal agreement, understanding, or provide answers. Both backward- and forward-looking functions can be present in the same utterance.

Determining the dialogue act corresponding to a given utterance is a non-trivial operation. The type of utterance only gives a partial indication of the underlying dialogue act – a question can for instance express a directive (“*Could you post this for me?*”). In order to accurately classify a dialogue act, a variety of linguistic factors have to be taken into account, such as prosody, lexical, syntactic and semantic features, and the preceding dialogue history (Jurafsky et al., 1998; Shriberg et al., 1998; Stolcke et al., 2000).

Dialogue acts are strongly contextual in nature: they are generally only intelligible within the particular conversational context in which they appear. Their successful interpretation must therefore venture beyond the boundaries of the isolated utterance. We briefly review here three striking examples of this subordination to context.

Deictic markers

A deictic marker is a linguistic reference to an entity that is determined by the context of enunciation. Examples of such markers are “*here*” (spatial reference), “*yesterday*” (temporal reference), “*this mug*” (demonstrative) or “*you*” (reference to an interlocutor). By their very definitions, deictic markers refer to different realities depending on the situation in which they are used: a “*here*” uttered in a classroom differs from a “*here*” uttered on the beach. As a consequence, the resolution of these expressions based on their context is a prerequisite to semantic interpretation.

Implicatures

As initially argued by Grice (1989), an important part of the semantics of dialogue acts is not explicitly stated but rather implied from the context. Consider the following constructed example:

A: Is William working today?
B: He has a cold.

In order to retrieve the “suggested” meaning behind B’s utterance – namely, that William is probably not working –, one needs to assume that B is cooperative and that his response is therefore relevant to A’s question. If an utterance initially seems to deliberately violate this principle, the listener must search for additional hypotheses required to make sense of the dialogue act. Grice (1989) formalised these ideas in terms of a cooperative principle composed of four conversational maxims that are assumed to hold in a natural conversation: the maxim of quality (“be truthful”), the maxim of quantity (“be exactly as informative as required”), the maxim of relation (“be relevant”) , and the maxim of manner (“be relevant”). These notions have been further developed by various theorists such as Wilson and Sperber (2002) and Horn and Ward (2008).

Non-sentential utterances

Non-sentential (also called elliptical) utterances are linguistic constructions that lack an overt predicate. They include expressions such as “*where?*”, “*at 8 o'clock*”, “*a bit less, thanks*” and “*brilliant!*”. Their interpretation generally requires access to the recent dialogue history to recover their intended meaning. This can lead to ambiguities in the resolution, as illustrated in these examples modified from Fernández et al. (2007):

- A: "When do they open the new station?" → B: "Tomorrow" (*short answer*)

- A: "They open the station today" → B: "Tomorrow" (*correction*)

- A: "They open the station tomorrow" → B: "Tomorrow" (*acknowledgement*)

2.1.3 Grounding

Dialogue acts are executed as part of a larger collaborative activity that requires the active coordination of all conversational partners, i.e. speaker(s) as well as hearer(s). This coordination takes place at various levels. The first and most visible level is the content of the conversational activity. The partners must ensure mutual understanding of each other's contribution, to control that they remain "on the same page". In addition, they also coordinate the process by which the conversational activity moves forward – by signalling that they attending to the partner who currently holds the conversational floor and acknowledging his/her contributions to the dialogue.

As an illustration, consider this short excerpt from a real conversation transcribed in the British National Corpus (Burnard, 2000) :[4]

DAVID: you know carriage a single carriage way?

CHRIS: mm

DAVID: now a ... a dual carriage way that has three lanes, what lane do you use?

CHRIS: the left ...

DAVID: right, what one you use for overtaking?

CHRIS: mm?

DAVID: what one you use for overtaking?

CHRIS: right

DAVID: middle lane

CHRIS: all three lanes is the taught now – middle?

DAVID: middle lane, you never stay in it either

CHRIS: so what's the last

DAVID: it's just

CHRIS: one for?

DAVID: turn right

CHRIS: turn, aye

We can observe in this short dialogue that the interlocutors constantly rely on the *common ground* of the interaction to move their discussion forward. They regularly check what pieces of information are mutually known and understood (e.g. "*you know carriage a single carriage way?*"). They also make use of a variety of signals to indicate when things are properly grounded ("*mm*", "*right*", "*turn, aye*") and when they are not ("*mm?*"). As the dialogue unfolds turn after turn, this common ground accumulates more and more material – for instance, the use of the middle lane for overtaking is not initially part of the shared knowledge for both speakers at the onset of the conversation, but becomes so towards the end.

---

[4]The full transcript is available at http://www.phon.ox.ac.uk/SpokenBNCdata/KBM.html

The common ground is defined as the collection of shared knowledge, beliefs and assumptions that is established during an interaction.[5] Each dialogue act is built upon the current common ground and participates in its gradual expansion and refinement. This process is called *grounding*. A variety of feedback mechanisms can be used to this effect. As described by Clark and Schaefer (1989), positive evidence of understanding can be expressed via cues such as:

Continued attention: The hearer shows that he/she continues to attend to the speaker;

Relevant next contribution: The hearer produces a relevant follow-up, as in the answer "*the left...*" following the question preceding it;

Acknowledgement: The hearer nods or utters a backchannel such as "*mm*", "*uh-uh*", "*yeah*", or an assessment such "*I see*", "*that's great*";

Demonstration: The hearer demonstrates evidence of understanding by reformulating or completing the speaker utterance;

Display: The hearer reuses part of the previous utterance, such as "*turn, aye*" after "*turn right*".

Communication problems can also occur, owing to e.g. misheard or misunderstood utterances. The hearer must in this case provide a negative feedback indicating a trouble in understanding. A large panel of clarification and repair strategies are available to recover from these communicative failures. These strategies include backchannels ("*mm?*"), confirmations ("*Do you mean that...?*"), requests for disambiguations, invitations to repeat, and tentative corrections.

All in all, these positive and negative signals enable the dialogue participants to dynamically synchronise what the speaker intends to express and what the hearers actually understand. This grounding process operates mostly automatically, without deliberate effort. It is closely related to the concept of interactive alignment that has recently been articulated by Garrod and Pickering (2004, 2009). Humans show a clear tendency to (unconsciously) imitate their conversational partners. In particular, they automatically align their choice of words, a phenomenon called lexical entrainment (Brennan and Clark, 1996). But alignment also occurs on several other levels such as grammatical constructions (Branigan et al., 2000), pronunciation (Pardo, 2006), accents and speech rate (Giles et al., 1991), and even gestures and facial expressions (Bavelas et al., 1986).

A proper treatment of grounding is critical for the development of intuitive conversational interfaces. As we already mentioned in the introduction to this thesis, understanding errors are indeed ubiquitous in spoken dialogue systems. The potential sources of misunderstandings are abundant, from error-prone speech recognition to out-of-domain utterances, unresolved ambiguities, and unexpected user behaviour. These pitfalls make the necessity of appropriate grounding strategies even more acute. Grounding for dialogue systems is an active area of research and important advances have been made regarding the formalisation of rich computational models of grounding (Traum, 1994; Matheson et al., 2000), the generation of clarification requests (Purver, 2004; Rieser and

---

[5]An information that is part of the common ground for a given group is more than simply known by every member of the group. They must also be aware that the information is shared and known by the other members. Formally speaking, a proposition p is part of the common knowledge for a group of agents G when all the agents in G know p, and they also all know that they all know p, and they all know that they all know that they all know p, *ad infinitum*. This definition can be rigorously formalised with set theory or epistemic logic (Meyer and Van Der Hoek, 2004).

Moore, 2005), the design of human-inspired error handling strategies (Skantze, 2007), the integration of non-verbal cues such as gaze, head nods and attentional focus (Nakano et al., 2003), and the development of incremental grounding mechanisms (Visser et al., 2012).

2.2 Spoken Dialogue Systems

2.2.1 Architectures

mention the question of adaptivity

2.2.2 Components

2.2.3 Applications

2.3 Dialogue Management

information state update

2.3.1 Hand-crafted approaches

Some topics investigated in this paradigm are the semantic and pragmatic interpretation of dialogue moves (Thomason and Stone, 2006; Ginzburg, 2012), the rhetorical structure of dialogue (Asher and Lascarides, 2005), or the use of plan-based reasoning to infer the user intentions (Allen and Perrault, 1980; Litman and Allen, 1987). These approaches are able to provide detailed analyses of various dialogue behaviours, but they generally assume complete observability of the dialogue context and provide only a very limited account (if any) of uncertainties.

(Grosz and Sidner, 1986) about plan-based approach.

Finite-state automata, frame-based, logical reasoning, etc.

2.3.2 Statistical approaches

This is typically done by representing the dialogue domain as a Markov Decision Process (MDP) or Partially Observable Markov Decision Process (POMDP) and subsequently estimating the parameters of these models from data (Lemon and Pietquin, 2007).

related to partial observability (noisy spoken inputs, unknown user intentions) and stochastic action effects (the user behaviour can be hard to predict).

Probabilistic modelling techniques must however face two important challenges. The most pressing issue is the paucity of appropriate data sets. Statistical models often require large amounts of training data to estimate their parameters. Unfortunately, real interaction data is scarce, expensive to acquire, and difficult to transfer from one domain to another. Moreover, many domains exhibit a rich internal structure with multiple tasks to perform, sophisticated user models, and a complex, dynamic context. In such settings, the dialogue system might need to track a large number of variables in the course of the interaction, which quickly leads to a combinatorial explosion of the state space.

MDP and POMDPs

2.4   Summary

5 points about dialogue? uncertain, structured, contextual, goal-driven, collaborative
    not said: predictions, incrementality, embodiment and situatedness, conversation structure

Chapter 3

Probabilistic Rules

Chapter 4

Learning from Wizard-of-Oz data

4.1 Bayesian parameter estimation

4.1.1 Key idea

4.1.2 Parameter priors

talk about simplifying assumptions: we are learning from partial data

4.1.3 Approximate inference

4.2 Estimation of action utilities from Wizard-of-Oz data

4.2.1 Data representation

4.2.2 Integrating the evidence

4.3 Experiments

4.3.1 Wizard-of-Oz data collection

4.3.2 Experimental setup

4.3.3 Empirical results

4.3.4 Analysis

4.4 Conclusion

Chapter 5

Learning from interactions

5.1 Bayesian Reinforcement Learning

5.1.1 Model-free methods

5.1.2 Model-based methods

5.2 Online planning

5.3 Experiments

5.3.1 Wizard-of-Oz data collection

5.3.2 User simulator

5.3.3 Experimental setup

5.3.4 Empirical results

5.3.5 Analysis

5.4 Conclusion

Chapter 6

User evaluation

Chapter 7

Concluding remarks

7.1   Summary of contributions

7.2   Future work

Formally characterise the expressivity of the rules and extend them to handle Ginzburg style update rules?

Appendix A

Relevant probability distributions

Uniform distribution

Multinomial distribution

Normal distribution

Dirichlet distribution

Kernel distribution

Should we include the last one?

Appendix B

Domain specification for HRI scenario

Appendix C

The openDial toolkit

Bibliography

J. F. Allen and C. R. Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178, 1980.

J. Allwood, J. Nivre, and E. Ahlsén. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9:1–26, 1992.

J. S Allwood. *Linguistic communication as action and cooperation : a study in pragmatics*. PhD thesis, Dept. of Linguistics, University of Göteborg, 1976.

N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, Cambridge, 2005.

J. L. Austin. *How to do things with words*. Harvard University Press, Cambridge, Mass., 1962.

J. B Bavelas, A. Black, C. R. Lemery, and J. Mullett. "i show how you feel": Motor mimicry as a communicative act. *Journal of Personality and Social Psychology*, 50(2):322–329, 1986.

H. P. Branigan, M. J. Pickering, and A. A. Cleland. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25, 2000.

Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology-Learning Memory and Cognition*, 22(6):1482–1493, 1996.

H. C. Bunt. Dynamic Interpretation and Dialogue Theory. In M. M. Taylor, F. Néel, and D. G. Bouwhuis, editors, *The Structure of Multimodal Dialogue, Volume 2*. John Benjamins, Amsterdam, 1996.

L. Burnard. User reference guide for the british national corpus. Technical report, Oxford University, 2000.

H. H. Clark. *Using Language*. Cambridge: Cambridge University Press, 1996.

H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive Science*, 13(2):259–294, 1989.

P. R. Cohen and C. R. Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212, 1979.

J. G. Core and J. F. Allen. Coding dialogs with the DAMSL annotation scheme. In *Proceedings of the Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, November 1997.

N. Dahlbäck, A. Jönsson, and Lars Ahrenberg. Wizard of oz studies: why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*, pages 193–200. ACM, 1993.

S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.

R. Fernández, J. Ginzburg, and S. Lappin. Classifying non-sentential utterances in dialogue: A machine learning approach. *Computational Linguistics*, 33(3):397–427, September 2007.

M. Frampton and O. Lemon. Recent research advances in reinforcement learning in spoken dialogue systems. *Knowledge Engineering Review*, 24(4):375–408, 2009.

R. M. Fung and K.-C. Chang. Weighing and integrating evidence for stochastic simulation in bayesian networks. In *Proceedings of the 5th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 209–220, 1989.

S. Garrod and M. J. Pickering. Joint action, interactive alignment, and dialog. *Topics in Cognitive Science*, 1(2):292–304, 2009. ISSN 1756-8765.

S. Garrod and M.J. Pickering. Why is conversation so easy? *Trends in Cognitive Sciences*, 8:8–11, 2004.

H. Giles, N. Coupland, and J. Coupland. 1. accommodation theory: Communication, context, and consequence. *Contexts of accommodation: Developments in applied sociolinguistics*, page 1, 1991.

J. Ginzburg. *The Interactive Stance*. Oxford University Press, New York, 2012.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, volume 1, pages 517–520 vol.1, 1992.

A. Gravano and J. Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601 – 634, 2011.

H.P. Grice. *Studies in the Way of Words*. Harvard University Press, 1989.

B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, July 1986. ISSN 0891-2017.

P. Heeman. Combining reinforcement learning with Information-State update rules. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 268–275, 2007.

Laurence Horn and Gergory Ward. *Handbook of pragmatics*, volume 26. Wiley-Blackwell, 2008.

E. Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '99, pages 159–166, New York, NY, USA, 1999. ACM. ISBN 0-201-48559-1.

D. Jurafsky, L. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical report, University of Colorado at Boulder Technical Report 97-02, 1997.

D. Jurafsky, E. Shriberg, B. Fox, and T. Curl. Lexical, prosodic, and syntactic cues for dialog acts. In *Proceedings of ACL/COLING-98 Workshop on Discourse Relations and Discourse Markers*, pages 114–120, 1998.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, May 1998.

S. Larsson. *Issue-based Dialogue Management*. PhD thesis, Gothenburg University, 2002.

S. Larsson, R. Cooper, E. Engdahl, and P. Ljunglöf. Information states and dialogue move engines. *Electronic Transactions on Artificial Intelligence*, 3(D):53–71, 1999.

O. Lemon and O. Pietquin. Machine Learning for Spoken Dialogue Systems. In *Proceedings of the 10th European Conference on Speech Communication and Technologies (Interspeech'07)*, pages 2685–2688, 2007.

S. C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

P. Lison. Declarative design of spoken dialogue systems with probabilistic rules. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2012)*, pages 97–106, 2012a.

P. Lison. Probabilistic dialogue models with prior domain knowledge. In *Proceedings of the SIGDIAL 2012 Conference*, pages 179–188, 2012b.

D. J. Litman and J. F. Allen. A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11(2):163–200, 1987.

C. Matheson, M. Poesio, and D. R. Traum. Modelling grounding and discourse obligations using update rules. In *Advances in Natural Language Processing*, pages 1–8, 2000.

M. F. McTear. *Spoken dialogue technology - toward the conversational user interface*. Springer, 2004.

J-J Ch Meyer and Wiebe Van Der Hoek. *Epistemic logic for AI and computer science*, volume 41. Cambridge University Press, 2004.

Y. I. Nakano, G. Reinstein, T. Stocky, and J. Cassell. Towards a model of face-to-face grounding. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 553–561, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

J. S. Pardo. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119:2382, 2006.

M. Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, King's College, University of London, August 2004.

A. Raux and M. Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *HLT-NAACL*, pages 629–637. The Association for Computational Linguistics, 2009. ISBN 978-1-932432-41-1.

V. Rieser and J. D. Moore. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 239–246. Association for Computational Linguistics, 2005.

N. Roy, J. Pineau, and S. Thrun. Spoken dialogue management using probabilistic reasoning. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 93–100. Association for Computational Linguistics, 2000.

S. J. Russell and P. Norvig. *Artificial Intelligence - A Modern Approach*. Pearson Education, 2010.

H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, dec 1974.

J. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969.

J. Searle. *Expression and Meaning. Studies in the Theory of Speech acts*. Cambridge University Press, 1979.

E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. Van Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492, 1998.

J. Sinclair and M. Coulthard. *Towards an Analysis of Discourse*. Oxford University Press, 1975.

G. Skantze. *Error Handling in Spoken Dialogue Systems: Managing Uncertainty, Grounding and Miscommunication*. PhD thesis, Royal Institute of Technology (KTH), Stockholm, 2007.

T. Stivers, N. J. Enfield, P Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P de Ruiter, K.-E. Yoon, and S. C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

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

N. Ström and S. Seneff. Intelligent barge-in in conversational systems. In *Proceedings of the 6th International Conference of Spoken Language Processing (ICSLP2000/Interspeech 2000)*, pages 652–655. ISCA, 2000.

R. H. Thomason and M. Stone. Enlightened update: A computational architecture for presupposition and other pragmatic phenomena. In *Presupposition Accommodation*. Ohio State Pragmatics Initiative, 2006.

M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28:675–691, 9 2005. ISSN 1469-1825.

D. R. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, Rochester, NY, USA, 1994.

D. R. Traum and J. F. Allen. Discourse obligations in dialogue processing. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 1–8. Association for Computational Linguistics, 1994.

D. R. Traum and E. A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8:575–599, 1992.

T. Visser, D. Traum, D. DeVault, and R. op den Akker. Toward a model for incremental grounding in spoken dialogue systems. In *The 12th International Conference on Intelligent Virtual Agents (IVA)*, Santa Cruz, CA, September 2012.

J. D. Williams. The best of both worlds: Unifying conventional dialog systems and POMDPs. In *International Conference on Speech and Language Processing (ICSLP 2008)*, Brisbane, Australia, 2008.

D. Wilson and D. Sperber. Relevance theory. *Handbook of pragmatics*, 2002.

S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24:150–174, 2010.

N. Lianwen Zhang and D. Poole. Exploiting causal independence in bayesian network inference. *Journal of Artificial Intelligence Research (JAIR)*, 5:301–328, 1996.

Index