1 Introduction

Amanda Stent and Srinivas Bangalore

The story of Pinocchio is the story of a man who creates a puppet with the ability to speak using human language, and to converse naturally – even to lie. Since the early days of the industrial revolution, people have imagined creating agents that can emulate human skills, including language production and conversation (Mayer, 1878; Wood, 2003). A modern example of this is Turing's test for artificial intelligence – a machine that is indistinguishable from a human being when evaluated through disembodied conversation (Turing, 1950).

This book concerns the intersection of two areas of computational research involved in the production of conversational agents – **natural language generation** and **interactive systems**.

1.1 Natural language generation

Natural language generation (NLG) systems are systems that produce human language artifacts (including speech, text, and language-rich multimedia presentations). NLG systems differ in the inputs they take, the types of output they produce and the degree to which they support interactivity. However, there are some common challenges across all types of NLG system. All end-to-end NLG systems perform the following three tasks: **content selection**, or determination of "what to say"; **surface realization**, or determination of "how to say it" (including assignment of content to media, selection of words, and arrangement of the content); and **production**, or the presentation/performance of the generated material (which in the case of an embodied agent may include production of body postures, gestures, and sign language). At the core of each NLG task is *choice*: for any decision, there may be numerous reasonable options, and the selection of a "right" option may depend on many factors including the intended recipient (the hearer or reader), the physical and linguistic context, and the persona of the agent itself. Consequently, NLG involves elements of creativity and subjectivity that are not typically thought of as key to computational tasks. This also complicates the evaluation of NLG systems, as there may be numerous reasonable outputs for a given input.

With respect to input, there are two broad categories of NLG system: **text-to-text** (which take text as input), and **data-to-text** (which start from non-linguistic inputs). There are interactive systems with NLG components of each type: for example, the MATCH dialogue system uses data-to-text generation to produce restaurant recommendations (Walker *et al*., 2002), while the Have2eat system uses text-to-text generation for the same purpose (Di Fabbrizio *et al*., 2010).

There are three broad categories of NLG system with respect to output: those that produce **language** only (text or speech), such as the COREF and PERSONAGE systems (Chapters 3 and 9); those that produce **multimedia** presentations, such as the ILEX (Cox *et al*., 1999) and interactiveX (Molina *et al*., 2011) systems; and those that generate the behaviors of **embodied conversational agents** (Cassell, 2000; André *et al*., 2001).

The standard pipeline for NLG systems since the mid-1990s, which reflects a Leveltian model of language production (Levelt, 1989; Reiter and Dale, 2000; Rambow *et al*., 2001), comprises four stages of processing corresponding to the three NLG tasks outlined earlier: content and discourse planning, sentence planning and surface realization, and production. While a strict pipeline architecture like this one may permit reuse of NLG components and simplify the implementation of each individual component, it also restricts the ability of NLG systems to be responsive and adaptive in very interactive contexts. This concern was expressed as early as 1995 (Kilger and Finkler, 1995), and recent years have seen a range of attempts to integrate the decision-making process across NLG tasks (for example, see Chapters 7 and 8).

1.2 Interactive systems

An interactive system is any computer system that engages in ongoing interaction with one or more human users. Although interaction can take many forms – an interactive photo album may only interact through images and gestures, a robot only through physical actions – in this book, we are concerned with systems that use human languages such as English, Korean, or sign language as a primary medium for interaction. Within this category, there is a wide range of types of system, both in terms of the amount and types of interaction supported and in terms of the flexibility and richness of the natural language produced by the system. At one extreme, there are systems that, while interactive, use very little natural language; an example is the baseline system for the recent GIVE challenges, which only said *warmer* and *colder* (Koller *et al*., 2010). Then, there are systems that use rich models of natural language, but are not very interactive, such as the BabyTalk system (Reiter *et al*., 2008). Finally, there are systems that are very interactive and use rich models of natural language, such as the ILEX museum guide (Cox *et al*., 1999). The contributions in this book contain numerous examples of language-intensive interactive systems, including instruction-giving agents for virtual environments (Chapter 8), embodied conversational agents (Chapter 10), and assistive and augmentative communication systems that facilitate communication between humans (Chapter 11).

1.3 Natural language generation for interactive systems

Consider the writing of a movie review for a magazine. The author has some content, which must be structured into a discourse. The author also probably has in mind the audience for the story, and targets the content of the review and the language used to that type of reader. However, the reader is (a) amorphous (a group of people rather than an individual); and (b) separated in time and space from the author (consequently, unable to reply or give feedback as the story is being constructed).

Now consider that a friend of the author asks about the movie over dinner. The author may communicate the same basic content, but everything else about the discourse is different. First of all, the discourse is **collaboratively constructed**; both people contribute to the language, topics, and overall shape of the conversation. This collaboration means that, at any point in the conversation, each person has their own mental model of the discourse. This introduces **uncertainty**: neither person knows if their mental model of the discourse matches that of the other person. Fortunately, the uncertainty can be ameliorated through **grounding** and **alignment**. Grounding behaviors (such as asking clarification questions, providing acknowledgments, and backchanneling) allow the conversational participants to check each other's models of the discourse. Alignment behaviors (such as converging on similar lexical or syntactic choices) allow the conversational participants to tailor their contributions to each other to minimize misunderstandings. Finally, because the author knows the friend, the content and form of the author's contributions to the conversation can be adjusted to make them particularly relevant and interesting for the friend, i.e., to encourage **engagement**. So we see that interaction affects every aspect of the NLG process. The chapters in this book address each of these aspects of interactive NLG.

1.3.1 Collaboration

In an interaction, no participant's contributions stand alone – the process of constructing the discourse is a **collaboration**. (This is true even if the participants are disagreeing or attempting to deceive each other – collaboration on an underlying task is not necessary for collaboration in the interaction process.) Each contribution is shaped by the need to fit into what has come before, and to maximize the chances of success of what will follow. These constraints influence content selection, surface realization and production. In addition, the collaboration causes additional language behaviours not present in non-interactive situations – these include interaction maintenance behaviours such as turn-taking.

The three chapters in Part I of the book highlight particular aspects of collaboration relevant to NLG for interactive systems.

When we imagine a collaboration, we tend to visualize something like bricklaying – where all participants are contributing to a common infrastructure (such as a brick wall) that is available to each participant. However, this is not typically the case for language-rich interactions. In a conversation, each participant keeps their own representation of their own internal state and of what they imagine to be the "shared state," and these representations may not be perfectly aligned (Clark, 1996). This introduces **uncertainty**. The management and reduction of uncertainty in conversation is carried out through processes including **grounding** and **alignment**. Grounding is the process through which conversational participants validate their mental models for the "shared state" of the interaction (Clark, 1996). Grounding behaviors include acknowledgments (e.g., *okay*), check questions (e.g., *you mean Sally from the store?*), and direct indications of understanding or of lack of understanding (e.g., *I didn't follow that*). Grounding behaviors are driven by an agent's understanding, but produced by NLG systems; the need to generate grounding behaviors affects the content selection, discourse planning, and surface realization processes. Alignment is a set of behaviors that conversational participants use to encourage and reinforce the development of similar mental models for the "shared state" of the interaction. For example, conversational participants may adopt the same postures, follow each others' eye gaze, repeat each others' gestures, adapt to each others' prosody and choice of words, and adopt each others' means of information packaging (such as a preference for longer or shorter contributions). The need to generate alignment behaviors primarily affects the sentence planning, surface realization, and production processes, but also means that conversational agents should maintain rich models of the linguistic content of the interaction.

The importance of modeling communicative intentions for advanced dialogue modeling is highlighted by Blaylock in Chapter 2. The chapter starts with an extensive survey of types of dialogue model – finite-state, slot-filling, and plan based – comparing and contrasting their ability to model communicative intentions. The author then introduces a **collaborative problem solving** model of dialogue which explicitly models communicative intentions. While most of the chapter is concerned with establishing the need for a representation of communicative intention, defining the range of communicative intentions and extracting them from user utterances, the chapter ends with some significant challenges at the intersection of dialogue management and content selection in dialogue systems.

In Chapter 3, Devault and Stone describe the importance of **grounding** for a successful conversation. They present a prototype system, COREF, that treats grounding as problem solving. COREF addresses both the issue of uncertainty in the system's understanding of the user's contributions, and the possibility of ambiguity in system-generated contributions. By maintaining a *horizon graph*, the system can evaluate the space of its possible contributions at each turn for implicatures and ambiguities, and select either a contribution that is acceptable under all interpretations, or a minimally underspecified contribution (in the case of ambiguity).

In Chapter 4, Purver *et al*. present an interesting phenomenon in human–human communication, **compound contributions**. Compound contributions are utterances spoken by a conversational participant that continue or complete an utterance spoken by another participant and result in a single syntactic and propositional unit distributed over multiple turns. The authors present interesting real-world examples drawn from the British National Corpus. They discuss the implications for modeling such contributions from a natural language generation perspective in human–machine dialogue. They sketch models of incremental joint parsing and generation using the grammatical formalism of **dynamic syntax** that can account for compound contributions.

Cambridge University Press
978-1-107-01002-4 - Natural Language Generation in Interactive Systems
Edited by Amanda Stent and Srinivas Bangalore
Excerpt
More information

1.3.2 Reference

A problem that is key to any interactive system is the understanding and production of **references** to entities' states, and activities in the world or in the conversation. Reference influences or is influenced by every dialogue task, from recognition of user input through production of user output. Referring expression generation, in particular, is the core of the NLG problem, for everything that a dialogue participant does is aimed at describing the world as it is or as the participant wishes it to be.

In the first chapter in Part II, Chapter 5, van Deemter introduces the reference problem and lays out a detailed argument for rich, comprehensive referring expression generation, with a series of increasingly comprehensive algorithms for generating referring expressions of increasing complexity. He starts by looking at the generation of simple referring expressions consisting of conjunctions of atomic properties by means of a **description by satellite sets** algorithm. He extends this algorithm to handle relations, other logical operators (including disjunction and negation), and different types of quantification. Through the discussion, the reader is introduced to modern knowledge representation techniques and to the question of the **expressive power** of referring expression generation algorithms.

In Chapter 6, Krahmer, Goudbeek, and Theune examine the interaction of choice and **alignment** during referring expression generation. In human–human conversation, a speaker uses their model of their listener(s) to generate referring expressions that uniquely identify the target object(s), as well as potentially serving other communicative functions. The challenge for the speaker is to construct brief, informative, and unambiguous referring expressions in a small amount of time. In this chapter the authors present a graph-based algorithm for referring expression generation. They demonstrate empirically that this approach can balance a speaker's preferences for referring expression construction and alignment constraints based on the history of the conversation.

1.3.3 Handling uncertainty

Grounding and alignment behaviors help minimize uncertainty related to each conversational participant's model of the interaction so far, but do not drive the interaction further. There is another type of uncertainty relevant to NLG, and that is the construction of a "right" contribution to drive the next step of the interaction. This topic is addressed in the two chapters in Part III of the book. These chapters present emerging approaches to NLG that model uncertainty in the dialogue state and that jointly optimise across different stages of the NLG pipeline.

As we said earlier, natural language generation has traditionally been modeled as a planning pipeline involving content selection, surface realization and presentation. Early NLG systems were rule-based, and consequently quickly grew complex to maintain and adapt to new domains and constraints. In Chapter 7, Lemon *et al.* present a reinforcement learning approach to NLG. The authors argue that the attraction of this new approach is that it relies on very little training data, is provably optimal in its decisions, and can quickly be adapted to new domains and additional constraints. They demonstrate how to use their approach to bootstrap NLG systems from small amounts of training data.

It has been widely acknowledged that the stages of the NLG pipeline are interdependent, even though they are typically modeled in isolation (primarily due to the combinatorial growth of the choices at each stage of the generation process). In Chapter 8, Dethlefs and Cuayáhuitl present an approach to jointly optimizing choices across multiple stages of NLG. They propose a reinforcement learning-based method that exploits a hierarchical decomposition of the search space in order to achieve computationally feasible solutions. They demonstrate the utility of their approach using the GIVE Challenge testbed.

1.3.4 Engagement

Any agent that produces language, whether for interactive or non-interactive contexts, will be more successful if it is **engaging**. An agent will engage a user if it is informative, relevant, clear, consistent, and concise (Grice, 1975). However, it is easier to be engaging for a particular (known) audience than for a general audience, and in a conversational interaction the participants are constantly giving each other information about themselves (what will best inform and engage them) and feedback about the success of the interaction so far. In particular, conversational participants give evidence of their culture and personality. In Part IV of the book we include three chapters that address the theme of **engagement** in NLG systems from diverse perspectives.

In the early 2000s, NLG researchers began to model parts of the language generation problem using **overgenerate and rank** techniques. Conceptually, these techniques use a (weak) model of language generation to produce a set of possible alternative realizations, and then rank or score the realizations using a language model. In Chapter 9, Mairesse employs this technique to address the issue of **personality-driven** surface realization. He shows that the effectiveness of interactive systems can be substantially influenced by the system's personality, as manifested in the linguistic style of system contributions. He presents a system that employs the overgenerate-and-rank technique as well as parametric predictive models for personality-driven surface realization. He also highlights the challenges in data-driven stylistic language generation techniques.

As communication aids become increasingly multimodal, combining speech with image and video input, so also natural language generation is evolving to produce multimodal contributions. One such multimodal generation mechanism is to use an **avatar** or a **virtual agent**. In the realm of multimodal communication, the intent of the message is carried not only by words, but also by the prosody of speech and by nonverbal means including body posture, eye gaze, and gestures. These paralinguistic means of communication take on a different dimension when contextualized by the cultural backgrounds of the conversational participants. In Chapter 10, the authors highlight the influence of cultural background in human communication and address the challenging problem of formulating and modeling culture for human-machine dialogues.

Another area in which engagement affects the performance of dialogue systems is when they are used to provide assistive and augmentative communication services. In Chapter 11, Tintarev *et al*. present an assistive and augmentative communication aid for helping children with language disabilities communicate their thoughts through language and multimodal interfaces. Here, the system must represent one user to other users during conversation, so must be adaptive and engaging. The chapter presents a system called “How was School Today?” that allows children to engage in a dialogue with their caregiver to discuss the events of their school day. The content planning component of the system summarizes information about where and when the child was during the school day. The sentence planning and surface realization components are fine-tuned to the application setting. The authors discuss possible future research aimed at exploiting richer contextual information to provide a more effective communication aid.

1.3.5 Evaluation and shared tasks

The last two chapters in Part V of this collection address the crucial issues of evaluation and shared tasks in NLG for interactive systems. The first chapter is a detailed discussion of the possibilities of eye tracking as an objective evaluation metric for evaluating the speech produced at the end of an NLG system, while the second is a sweeping survey of shared tasks and evaluation in NLG and interactive systems.

With the increasingly multimodal nature of human–machine conversation, where humans interact with devices endowed with cameras, the efficacy of conversation can be hugely improved by taking into account different factors beyond just the speech input/output to/from the system. In Chapter 12, White *et al*. present a detailed analysis of the impact of prosody on speech synthesis. They propose to include the eye tracking movements of humans listening to synthesized speech as an additional metric to evaluate the prosody of speech synthesis. Their experimental results indicate that the fundamental frequency (F_0) feature of the synthesised speech plays an important role in the comprehension of the speech, particularly in a contrastive discourse context.

Since the early days, NLG researchers have embedded their components in real-world applications. However, since these applications are tightly tied to particular tasks, they do not permit comparison of different NLG techniques. In Chapter 13, Belz and Hastie provide a comprehensive study of recent efforts to provide evaluation frameworks and shared evaluation campaigns for a variety of NLG tasks, including surface realization, referring expression generation, and content planning. These campaigns allow the NLG community to compare the effectiveness of different techniques on shared data sets, thus decoupling NLG components from the context of an overall application. It also helps to lower the barrier for entry to the NLG field. However, as Belz and Hastie point out, NLG evaluation in interactive systems has not been a focus in these efforts. Given that conversation is collaborative, it is challenging to evaluate the success of an individual system contribution, and especially to separate task success (over a whole conversation) from linguistic success (of any individual NLG decision). This research direction in NLG evaluation is ripe for further exploration.

1.4 Summary

The contributors to this collection intend to spur further research on NLG for interactive systems. To this end, each chapter contains concrete ideas for research topics. The book is also accompanied by a website, www.nlgininteraction.com, containing pointers to software and data resources as well as a bibliography.

References

André, E., Rist, T., and Baldes, S. (2001). From simulated dialogues to interactive performances. In *Multi-Agent-Systems and Applications II*, pages 107–118. Springer, Berlin, Germany.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, **43**(4):70–78.

Clark, H. (1996). *Using Language*. Cambridge University Press, Cambridge, UK.

Cox, R., O'Donnell, M., and Oberlander, J. (1999). Dynamic versus static hypermedia in museum education: An evaluation of ILEX, the intelligent labelling explorer. In *Proceedings of the Conference on Artificial Intelligence in Education*, pages 181–188, Le Mans, France. International Artificial Intelligence in Education Society.

Di Fabbrizio, G., Gupta, N., Besana, S., and Mani, P. (2010). Have2eat: A restaurant finder with review summarization for mobile phones. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 17–20, Beijing, China. International Committee on Computational Linguistics.

Grice, H. P. (1975). Logic and conversations. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics III: Speech Acts*, pages 41–58. Academic Press, New York, NY.

Kilger, A. and Finkler, W. (1995). Incremental generation for real-time applications. Technical Report DFKI Report RR-95-11, German Research Center for Artificial Intelligence – DFKI GmbH, Saarbrücken, Germany.

Koller, A., Striegnitz, K., Byron, D., Cassell, J., Dale, R., and Moore, J. D. (2010). The first challenge on generating instructions in virtual environments. In Krahmer, E. and Theune, M., editors, *Empirical Methods in Natural Language Generation*, pages 328–352. Springer LNCS, Berlin, Germany.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Mayer, A. M. (1878). On Edison's talking-machine. *Popular Science Monthly*, **12**:719–723.

Molina, M., Parodi, E., and Stent, A. (2011). Using the journalistic metaphor to design user interfaces that explain sensor data. In *Proceedings of INTERACT*, pages 636–643, Lisbon, Portugal. Springer.

Rambow, O., Bangalore, S., and Walker, M. (2001). Natural language generation in dialog systems. In *Proceedings of the Human Language Technology Conference (HLT)*, pages 270–273, San Diego, CA. Association for Computational Linguistics.

Reiter, E. and Dale, R. (2000). *Building Natural Language Generation Systems*. Cambridge University Press, Cambridge, UK.

Reiter, E., Gatt, A., Portet, F., and van der Meulen, M. (2008). The importance of narrative and other lessons from an evaluation of an NLG system that summarises clinical data. In *Proceedings of the International Conference on Natural Language Generation (INLG)*, pages 147–156, Salt Fork, OH. Association for Computational Linguistics.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, **59**:433–460.

Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J. D., Johnston, M., and Vasireddy, G. (2002). Speech-Plans: Generating evaluative responses in spoken dialogue. In *Proceedings of the International Conference on Natural Language Generation (INLG)*. Association for Computational Linguistics.

Wood, G. (2003). *Edison's Eve: A magical history of the quest for mechanical life*. Anchor Books, New York, NY.