Question Generation in the CODA project

Paul Piwek and Svetlana Stoyanchev

Centre for Research in Computing,
The Open University, Milton Keynes, UK
{p.piwek,s.stoyanchev}@open.ac.uk

Abstract. In the ongoing CODA project, we are developing a system for automatically converting monologue into dialogue. The dialogue is generated in a two-step approach. Firstly, snippets of input monologue are mapped to dialogue act sequences. Secondly, these sequences are verbalized. The conversion relies partly on analysing input monologue in terms of its discourse relations. This short paper briefly describes the approach to the first step in CODA. This approach involves the use of a parallel corpus of monologues and dialogues to learn mappings from monologue to dialogue acts. Here, we focus on dialogue acts that involve question asking.

Key words: dialogue generation, parallel monologue/dialogue corpus, language generation, question generation

1 Introduction

CODA[1] is a 2-year project that started in July 2009. The overall aim of CODA is to develop a system for automatically generating dialogue from monologue. This paper provides a brief overview of the work so far, concentrating on the generation of questions (which is a part of generating dialogue).

Dialogue as a means of information presentation goes back at least as far as the ancient Greeks – Plato conveyed his philosophy through fictitious conversations between Socrates and his contemporaries. In this kind of dialogue, henceforth *expository dialogue*, the main purpose is to inform the reader or audience; the information the dialogue partners convey to one another is subservient to this purpose.

Craig et al. [2] found that presenting tutorials as dialogues that are viewed by the student has advantages over presenting the same information as monologue; the benefits of dialogue include stimulating students to write more in a free recall test and to ask twice as many deep-level reasoning questions in a transfer task. Further benefits of dialogue presentation are summarized in [8].

Despite the evidence that in some situations dialogue can be more effective for presenting information than monologue, for the foreseeable future, *text* in monologue form is likely to remain the most common medium for the production of information content, e.g., in the form of books, articles, web pages, and leaflets.

[1] COherent Dialogue Automatically generated from text

Writing monologue is practised by most authors while writing in the form of a dialogue is a less common practice. The aim of CODA is to allow existing monologue content to be presented as dialogue. In this project we develop the theory and technology for automatic transformation of text in monologue form to expository dialogue, specifically dialogues between a 'layman' (e.g., patient or student) and 'expert' (e.g., doctor or tutor) character. Our goal is to present monologue in the form of a dialogue to observers with the purpose of informing them. It is different from the task of interactive tutoring dialogue systems [4, 3, 12] where the system plays a role of a tutor and a user is one of the participants in the dialogue.

The approach to dialogue generation taken in CODA differs from most of the existing work on expository dialogue generation, e.g., for Embodied Conversational Agents (see [11]), in that it starts from text rather than information stored in a database or knowledge representation. The work also differs from the pilot study we carried out on text to dialogue generation (see [7]), in that it aims to use mapping rules that have an empirical grounding. In particular, we have constructed a corpus by translating professionally authored dialogues into monologues. The CODA parallel corpus[2] consists of dialogue segments aligned with monologue snippets expressing the same content. A recent study showed that human-authored monologue-to-dialogue mapping rules can account for only a small percentage (about 13%) of the dialogue structures that are found in professionally authored dialogues [5].

In the next section, we introduce the overall architecture of the CODA system and the corpus from which monologue to dialogue mapping rules were extracted. In Section 3, we describe our findings relating to the generation of questions. The paper ends with a conclusions section.

2 Generating Dialogue in CODA

For the CODA system, see Figure 1, the input is a monologue. Off-the-shelf technologies are used for parsing its syntactic and discourse structure. The annotated text is processed in two steps. Firstly, the text is mapped to a sequence of dialogue acts. In a second step, these acts are verbalized. The resulting dialogue can be rendered either as text (via 6a and 7a) or as a video of computer-animated agents (6b and 7b), with MPML3D [9] as the interface language to the agents.

The repository of rules for mapping discourse relations in text to dialogue acts is automatically extracted (for more detail see [8]) from the CODA corpus. The corpus currently consists of 800 turns from samples of dialogues written by acclaimed authors such as Mark Twain. For each dialogue sample, a monologue was written expressing the same content. Dialogue spans were then aligned with snippets of monologue. The dialogue side was annotated with dialogue acts and the monologue side with discourse relations. The corpus construction is described in [10] together with interannotor agreement information.

[2] The corpus will be released at `http://computing.open.ac.uk/coda/` in June 2010.

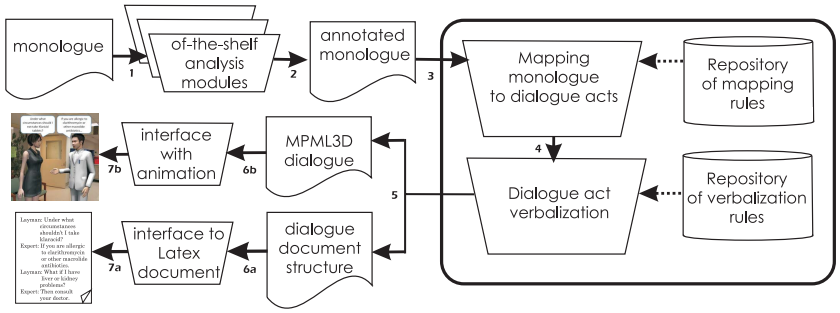

Fig. 1. CODA Architecture for transforming text to dialogue

3 Questions in Dialogue

A dialogue is a sequence of dialogue acts including questions, answers, and responses signalling agreement or contradiction. The CODA corpus aligns dialogue sequences written by professional writers with monologue segments. In our corpus we use three types of question dialogue acts (presented together with examples and frequency from a 259 turn sample of the CODA corpus):

Question type	*Example*	*Frequency*
Yes/No	'Can it force a right-principled man to do a wrong thing?'	173 instances
Complex	'How are you going to make that out, when the lower animals have no mental quality but instinct, while man possesses reason?'	65 instances
Factoid	'How many times did you try the experiment?'	18 instances

Table 1. Types of question asking dialogue acts in the CODA corpus and their frequencies in a 259 turn sample

Dialogue (including questions) generation in CODA project involves two steps:

1. determining the dialogue act sequence; and
2. verbalizing the dialogue sequence.

Our mapping rules from text to dialogue acts address the first step of dialogue generation: determining which dialogue acts may be used to express a particular bit of input monologue in dialogue generation. This task is closely related to the Nielsen's Question Type Determination task [6]. Nielsen's task, one of three tasks central to Question Generation, is about determining the most appropriate type of question which can be asked for an input text. Our task, in addition to question type determination, also involves identifying dialogue acts for the responses to the questions, such as *Agree* or *Contradict*. To illustrate this point, let us look at a specific rule that was automatically extracted from the CODA corpus:

$$\textsc{attribution}(P,Q) \quad \Longrightarrow \quad \begin{array}{l}\text{Expert: yes/no InfoRequest}(Q),\\ \text{Layman: Resp-Answer-Yes}(P)\end{array}$$

This rule was extracted from the following dialogue fragment in the CODA corpus (out of Twain's *What is man?*):

Expert: He felt well?
Layman: One cannot doubt it.

In the corpus, the dialogue fragment is aligned with the following monologue segment:

[One cannot doubt]$_1$ [that he felt well]$_2$'

The monologue segment is accompanied by an annotation with the discourse relation Attribution(1,2).

70% of the rules that we automatically extracted are based on a discourse relation in the monologue. In other words, most sequences of dialogue acts (forming a coherent dialogue fragment)[3] were aligned with text held together by discourse relations. However, not all rules involve discourse relations. Some monologue snippets map to several dialogue acts, even though they do not include a discourse relation. For example, the monologue snippet

> However, on his way home his mind was in a state of joy which only the self-sacrificer knows.

maps to the following dialogue fragment:

Expert: What was his state of mind on his way home?
Layman: It was a state of joy which only the self-sacrificer knows.

In this example, the *what* question is based on the semantics of the statement, not on discourse structure.

Both for rules with and without discourse relations the majority (74% and 78%, respectively) mapped to dialogue act sequences that included a question.

The second task of dialogue sequence verbalization is achieved with dialogue move verbalization rules. Such rules for verbalizing questions have been constructed in previous work on question generation (e.g., Wyse and Piwek [13]). In our dialogue generation task, we reuse previously constructed question generation rules as well as rules that are harvested from the CODA corpus. The corpus provides sentence-question pairs from which we aim to extract, in the first instance, manually further dialogue move verbalization rules. We aim also aim to experiment with methods from the paraphrasing literature (e.g., Barzilay and McKeown [1]) for automatically deriving the dialogue act verbalization rules.

[3] I.e., translatable to a sentence or a paragraph that conveys a complete point.

4 Concluding Remarks

This paper described the ongoing CODA project and the generation of questions as an integral part of dialogue generation from monologue. We have provided some information on questions found in the parallel CODA corpus and the relation of these questions to the monologue with which they are aligned in the corpus.

We are still early on in the project, but have already several outcomes. These include a parallel corpus of monologues and dialogues (and a dedicated tool for creating such corpora) and a repository of high-level rules for mapping discourse relations to dialogue act sequences. We're currently working on verbalization of the dialogue acts. For the second year of the project an evaluation study is planned. This will focus on evaluating the fluency and coherence of automatically generated dialogue, and also whether the generated dialogues preserve the information in the input monologues.

Finally, with regards to the CODA corpus, we believe that it can be a useful resource not just for monologue-to-dialogue generation, but also more generally for question generation.

Acknowledgments. We would like to thank the three anonymous reviewers for QG2010 for their helpful comments. The research reported in this paper is funded by the UK Engineering and Physical Sciences Research Council under grant EP/G/020981/1.

References

1. R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *Proc. of ACL/EACL*, Toulouse, 2001.
2. S. Craig, B. Gholson, M. Ventura, A. Graesser, and the Tutoring Research Group. Overhearing dialogues and monologues in virtual tutoring sessions. *International Journal of Artificial Intelligence in Education*, 11:242–253, 2000.
3. Martha W. Evens and Joel A. Michael. *One-on-one Tutoring by Humans and Machines*. Mahwah, NJ: Lawrence Erlbaum Associates, 2006.
4. Arthur C. Graesser, S. Lu, G.T. Jackson, H. Mitchell, M. Ventura, A. Olney, and M.M. Louwerse. AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers*, 36:180–193, 2004.
5. N. Hastings. Towards Transforming Monologues into Dialogues Automatically. Master's thesis, Computing Department, The Open University, March 2010.
6. R. Nielsen. Question generation: Proposed challenge tasks and their evaluation. In V. Rus and A. Graesser, editors, *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, Virginia, September 2008.
7. P. Piwek, H. Hernault, H. Prendinger, and M. Ishizuka. T2D: Generating Dialogues between Virtual Agents Automatically from Text. In *Intelligent Virtual Agents*, LNAI 4722, pages 161–174. Springer Verlag, 2007.

8. P. Piwek and S. Stoyanchev. Generating Expository Dialogue from Monologue: Motivation, Corpus and Preliminary Rules. In *Procs of NAACL-HLT 2010*, Los Angeles, June 2010.
9. H. Prendinger, S. Ullrich, A. Nakasone, and M. Ishizuka. MPML3D: Scripting Agents for the 3D Internet. *IEEE Trans. on Visualization and Computer Graphics*, 2010.
10. S. Stoyanchev and P. Piwek. Constructing the CODA corpus. In *Procs of LREC 2010*, Malta, May 2010.
11. K. van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann. Fully generated scripted dialogue for embodied agents. *Artificial Intelligence Journal*, 172(10):1219–1244, 2008.
12. Kurt VanLehn, Arthur C. Graesser, G. Tanner Jackson, Pamela W. Jordan, Andrew Olney, and Carolyn P. Rosé. When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1):3–62, 2007.
13. B. Wyse and P. Piwek. Generating questions from openlearn study units. 2009.