Planning Transition Relevance Points in Speech-based Information Systems

Yan Qu
Computational Linguistics Program
Carnegie Mellon University
Pittsburgh, PA 15213
yqu@cs.cmu.edu

Abstract

The ability for a system to recognize when to take initiative and when to let the other party take the initiative is essential to the design of an effective human-computer dialogue system. However, there are not yet general guidelines on where a system should place the turn transition relevance points (TRPs). In this paper, we report some observations on the relationship between TRPs and the amount of information presented in human dialogues and in existing dialogue systems. The feasibility of planning TRPs based on the measure of information is investigated in a user study. Preliminary results of the user study are discussed.

Introduction

Findings by conversation analysts on analysis of human-human conversations demonstrate that although there are frequent periods of overlap in which more than one conversant is speaking, these periods are usually brief. Humans seem to have an economic way of dividing "floor" control among the conversants, alternating their turns as the performing speaker. Conversation analysts define a *transition relevance point* (TRP), where the turn could switch between parties. At the TRP point, a simple rule system can be used to determine who would be expected to speak next (Sacks, Schegloff, & Jefferson 1978).

Planning TRPs is a very important design issue in developing mixed initiative human-computer interaction. The ability for the system to recognize when to take initiative and when to let the other party take the initiative is essential to the design of an effective interactive system. This ability is even more important for speech-based human-computer dialogue systems. A response of paragraph length may be acceptable in screen-based human-computer interaction; however, such a response in speech, e.g., over the telephone, might be very inefficient: the user probably will have a hard time in remembering the information given by the system without asking the system to repeat part or all of the response.

Research has been done in the studies of discourse structures to identify the distinguishing features at TRPs by which speakers signal information about discourse structure segments to listeners: cue words (Grosz & Sidner 1986; Reichman 1985), intonation (Hirschberg & Pierrehumbert 1986), boundary tones and silences (Traum & Heeman 1996), control phrases and topic (Whittaker & Stenton 1988) and pronominalization (Guindon *et al.* 1986; Brennan, Friedman, & Pollard 1987). These findings, however, do not offer us a general theory of discourse, because the cues that can be used as indicators of a particular type of shift may occur infrequently. Moreover, these approaches identify the signaling features when a turn shift occurs, but leave unspecified why a transition is placed at one point, and when TRPs should be generated.

Planning the TRPs by computer systems is closely related to the problem of content planning. An important research question in content planning is how much information a system should generate when giving a response. For example, in (Moore & Paris 1992; McKeown 1985; Maybury 1991), the system responses to the users' questions are typically a few utterances long. It is impossible for the user to signal misunderstanding in the middle of a system response. This makes it harder for the user to raise questions at relevant positions of information giving or explanation, and harder for the system to guess the source in case of trouble. Haller (Haller 1994; 1995) develops on Moore's work by separating two types of text plans: discourse content plans and content-selection plans. To achieve discourse goals, her system uses the discourse text plans selecting some minimal text content. Upon users' follow up questions, the Interactive Discourse Planner can augment this content with additional content by using one or more content-selection text plans.

An implicit assumption underlying the work of text planning mentioned above is that the TRPs are placed after explanations or some minimal text plans. Planning of TRPs is not explicitly considered during content planning. A noticeable exception is Cawsey's work on generating explanatory dialogues (Cawsey 1992). Cawsey explicitly plans TRPs within the text plans, allowing the user to ask clarification questions and make acknowledgements, and allowing the system to restruc-

ture and prune the plan yet to be executed. However, Cawsey does not specify the basis for planning the TRPs.

In this paper, we describe an approach of using usability study to optimize the placement of TRPs. Many factors can contribute to where a TRP should be placed. We focus in this paper on the relationship between the placement of TRPs and the amount of the information presented. We first report some observations on the relationship between TRPs and the amount of information presented in human dialogues and in existing dialogue systems. We then report a user study which aims to determine the feasibility of planning TRPs based on the amount of information. Preliminary results of the user study are discussed.

Information and TRPs

We measured the amount of information in a presentation segment as the number of content words in that segment. Human information-giving dialogues and human-computer dialogues were analyzed in order to get the relationship between the amount of information and the placement of TRPs. The dialogues we analyzed are summarized in Table 1.

The human-human information dialogues are push-to-talk dialogues, in which the speakers signal the change of turn by pushing a button. The reason to choose push-to-talk dialogues is that in push-to-talk dialogues where to switch a turn is controlled by the speaker, and thus there is a need for the speaker to plan when to give the turn to the other speaker. We exclude cases of repairs, so that we can assume complete transfer of information.

Figure 1 shows the average amount of information in a presentation segment in the four types of human information-giving dialogues (hotel, flight, ground transportation, and service) and the four human-computer dialogues systems (Abella, Brown, & Buntschuh 1996; Cawsey 1992; McKeown 1985; Maybury 1991). The four human-computer dialogue systems were chosen for analysis because of the number of examples available. We only consider the information giving part of dialogues in these systems. We have the following observations:

- In the human dialogues, the average amount of information presented in a segment is around 10 content words.
- The two screen-based systems (McKeown 1985; Maybury 1991) provide much more information in one presentation segment than that in human conversation.
- The speech-based system (Abella, Brown, & Buntschuh 1996) provides much less information in one presentation segment than that in human conversation.
- The average amount of information in one presentation segment in Cawsey's system falls within the range of that of human conversation. However, it is unclear how she determines where to plan TRPs,

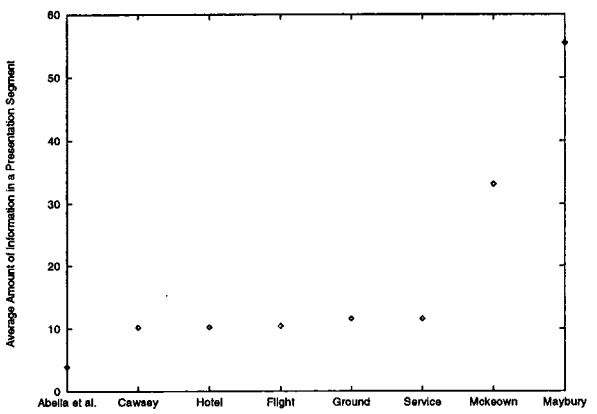

Figure 1: Average Amount of Information in a Presentation Segment Across Systems

Pilot Study: TRPs and Usability

The analysis of the human-human dialogues and the human-computer dialogues in the previous section shows that without a consideration of how much information to present in one presentation segment, a system could present too much or too little information with respect to human cognitive capacity. User models are usually proposed as the solution to controlling the amount of information. However, detailed user models are generally not available.

We propose to use the amount of information as the basis for optimizing the size of presentation and the placement of TRPs. The effectiveness of this measure was tested in a user study. The method is based on the assumption that where the system should generate TRPs is a system design issue, and that, similar to other design issues, the placement of the TRPs can be considered as a system design parameter that may vary. Thus we can use system design evaluation techniques to evaluate a particular setup and to get to the optimal setup based on system evaluations. As the ultimate use of a human-computer interaction system is to help user in obtaining information or problem solving, we measure the system performance based on the usability of the system. Specifically, the procedure is as follows:

1. specify different settings for TRP placement;
2. get the corresponding usability measure for each setting of TRP placement;
3. find out the functional dependency between system usability and the TRP placement;
4. find the optimal TRP placement.

Since any usability study of an interactive system would require the system's ability to generate TRPs

Human Information Dialogues:	
Hotel	Hotel information dialogues
Flight	Flight information dialogues
Service	Service information (e.g., banks, post offices) dialogues
Ground	Ground transportation information dialogues
Human-Computer Dialogues:	
Abella, Brown, and Buntschuh	Flight reservation dialogues
Cawsey	Explanation dialogues
Mckeown	Military database information dialogues
Maybury	Military database information dialogues

Table 1: Dialogue Corpus

in the first place, we propose to start with a usability study in a non-interactive setting. The findings from the non-interactive system could give hints on the placement of the TRPs in an interactive system. The optimal non-interactive system could be used as the baseline performance measure for future systems with added interactivity.

Experiment

Speech-interface information systems

Speech-interface information systems are of research interest for several reasons: First, there are situations where speech is the only mode of communication, such as information access through telephone. Second, speech-interface systems provide an appropriate interface for tasks such as eyes-busy/hands-busy tasks where verbal interaction is the most effective medium of communication, as in the case of automobile navigation. Thirdly, speech-based information services can provide convenient access to information using relatively simple interface devices — such as the telephone — and using a medium that is familiar and simple to use — such as speech.

Application and task domain

The application domain for the current experiment is information giving. Practical applications of such a domain include message centers and database query systems. Information giving is also an integral part of problem-solving dialogues, where distributed information needs to be communicated. However, information giving is different from the problem-solving task in that in information giving, once the query from the user is understood by the system, the system is usually capable of producing a response based on the database knowledge and the available user model. Given the planned response, yet another important decision for the system to make is in what way to present the information so that the user can get the information effectively. This is a decision necessary for the system regardless of interactivity.

The task domain for our experiment is hotel information. The hotel information domain involves an intermediate level of complexity in both information content and interaction with the user. Information content is potentially complex because of the large number of objects and object types in the domain (prices, type of rooms, restaurants, services, addresses, transportation, availability etc). Interaction with the user is potentially complex because both the specification of queries and the delivery of information may require multiple, interactively controlled steps, and because the user is provided with the flexibility to pursue these steps in different orders, to give multiple inputs at one time, to request repetitions and clarifications, or to start over at any point in the dialogue. Since in this paper, we focus on information giving in non-interactive systems, we do not consider interaction issues any further.

System setup

The current database contains information about five hotels in Pittsburgh. For each hotel, information such as the kinds of rooms, prices, availability, restaurants, conference rooms, and parking is available. Given a query about a particular hotel, the system generates a description of the hotel based on the information available in the database, and produces a description through speech. The speech is produced by the DECtalk software running on a DEC Alpha machine. The speech is produced at 180 wpm (words per minute), the default rate of the DECtalk software. The lengths of the descriptions range from 100 words to around 170 words. The average number of words is 128 per description.

The placement of TRPs within the multi-sentence description is dynamically planned based on the amount of information specified by the designer. For example, if the average amount of information per presentation, or the number of content words in a segment, is set at 20, then the generated system responses will be broken in presentation segments, with each segment containing an average of that amount of information, and TRPs will be placed at those points. In the experiment, we tested five system setups. In one setup, the TRPs were set with an average amount of information of 50. For the description in Figure 2, this corresponds to setting the TRP at the end of the entire description. In another setup (Figure 3), the TRPs were placed with

an average amount of information at 5, which roughly corresponds to planning a TRP after every sentence in the description. The other three setups were between these two extremes, with the TRPs set with information size 10, 20, and 40. Each TRP was delimited by a longer pause than the normal pauses that are not TRPs.

> Welcome to the Best Western Hotel. The hotel has single and double bed rooms. Both kinds of rooms are fifty dollars per person per night. Continental breakfast and maid service are included in the price. The shuttle service goes from the airport right outside the baggage claim, and it will drive directly to the hotel. It runs every hour on the half hour starting at nine in the morning. For payment, the hotel accepts visa, mastercard, traveller's checks, and cash. The address of the hotel is 875 Greentree Road. The contact phone number is 699 2905. <TRP>

Figure 2: Information Giving with TRP at the End

> Welcome to the Holiday Inn. <TRP> The Holiday Inn has rooms with two double beds or one king-sized bed. <TRP> Both kinds of rooms are one thirty dollars per night. <TRP> The holiday Inn has a restaurant. <TRP> Continental breakfast is free for kids twelve and under. <TRP> The hotel is near Oakland, which has a lot of inexpensive restaurants. <TRP> Room service is available. <TRP> There is adequate parking at the hotel. <TRP> Shuttle service is available to and from the airport. <TRP> All major credit cards, travellers checks, cashier's checks and cash are accepted. <TRP> The address of the hotel is 401 Oakland Avenue. <TRP> The contact phone number is 661-1476. <TRP>

Figure 3: Information Giving with TRP after Each Sentence

Data collection

The information for each hotel in the database was randomly assigned a particular TRP setting. Six subjects participated in the usability study. Each subject listened through all five hotel information descriptions with their respective system setups. A subject went through these setups in a randomized order so that learning effects can be avoided. The subjects were briefed about the system, and were asked to obtain information for a particular hotel from the system in a particular setup. The query about a particular hotel was pre-defined in the instructions, so the subjects did not need to interact with the system. The subjects

```
room type:
price:
room service:
parking:
payment:
address:
contact phone:
```

Figure 4: An Example Template

were asked to first select the features they were interested in from a given list of hotel features, and then to record the information for those chosen features by filling out template slots. Currently, the given list of features consists of 17 hotel features, including room types, price, restaurants, services, and parking. Figure 4 lists some features appearing in the template. Not all features are addressed in the system descriptions since related information is not always available from the database. The most common features chosen by the subjects are room types, price, location, breakfast, TVs, and address. Only exact matches filled out for features were counted as correct answers. If a feature has more than one value, such as price, each value is counted separately. A practice session was conducted to familiarize the subjects with the system and the template-filling task. After the experiment, the subjects were asked to fill in a questionnaire, which solicited their impression of the system setups, limitations of the system, and recommendations for improvements.

Usability and efficiency measures

Usability of the system is measured quantitatively using primarily the recall and precision of one or more system runs. The definitions of precision and recall are similar to those found in the area of message understanding (Grishman & Sundheim 1996):

$$recall = \frac{N_{correct}}{N_{key}} \quad (1)$$

$$precision = \frac{N_{correct}}{N_{correct} + N_{incorrect}} \quad (2)$$

in which the answer key has N_{key} filled slots, and that the user fills $N_{correct}$ slots correctly and $N_{incorrect}$ incorrectly. The users can request the system to play back the descriptions. It is hypothesized that when the experiment stops, the user can get near perfect recall and precision for the information recorded.

The efficiency of the system is measured by the total number of system runs and the total time needed before the user stops. The total number of runs is decided by the users, when they feel their request for information is satisfied. The total amount of time needed is the sum of the time taken for system to utter the description and the time for pauses (including the normal

pauses after commas and periods, and the extended pauses inserted after TRPs).

Preliminary Results

Table 2 summarizes the results of the user study. The values of the TRP settings indicate the average amount of information, or the number of content words, in a presentation segment. For a particular TRP setting, a user may request more than one system run. The time recorded for a system run is the sum of speech time and pause time. The speech time is the actual time to utter texts. The pause time includes the normal pause time for commas (one sixth of a second), periods (half second), and the extended pauses for TRPs (4.5 seconds). The recorded times for a given setting are cumulative with increasing system runs. Since the descriptions for different hotels vary in length, the recorded times are normalized to a rate of per 100 words. The normalized times are reported in Table 2 in seconds. Precision and recall for a given system run are the average precision and recall of all the users for that system run. Not all the users require the same number of system runs. If, for a particular system run, some users do not participate, the precision and recall values of their last system runs will be used for calculating the averages of the current one.

The following is a summary of the results reported in Table 2:

- The precision for each system run is generally high for every TRP setup. We observed that the users only recorded information when they had certain confidence.
- Recall generally increases with system runs. Note that setup 40 is an exception. In this setup, most of the subjects had difficulty recording the address of the hotel in the description. After the third system run, they either wrote down wrong addresses or left the address slot blank.
- Users do not always get perfect precision and recall. In some cases, the users falsely believed that they got all the correct information. Another cause of the imperfect precision and recall is that the users had problems getting some names. In the latter case, simply placing TRPs with extended pauses will not solve the problem. More complicated behavior, such as interaction with the user, is needed of the system.
- There is no noticeable performance difference between presentation segments with average amount information of 5 and 10, yet setup 10 is slightly more efficient than setup 5 in terms of the system time per 100 words. This suggests that within certain ranges, finer TRP placement does not necessarily improve system usability.
- Even though more TRPs increase the time for a single system run, system setups with more TRPs actually take less total time to get to certain performance levels, e.g., 90% in recall, than those with fewer TRPs. For example, in Table 2, the system time per 100 words decreases when the amount of information in a presentation segment increases for corresponding system runs. However, to get to the same level of recall, the users requested more system runs in setups with fewer TRPs, thus making the total system time per 100 words larger. Note again that setup 40 is an exception. Because of the difficult names in this description, the performance in this setup is worse than that in setup 50.

The results of the user study suggest setup 10 to be a good design setting both in terms of performance and efficiency. This setting is also supported by the empirical observations of human information-giving dialogues discussed previously. However, because of the limited scope of the user study, the reported results are only preliminary and need further investigation.

It should be pointed out that planning TRPs simply based on the amount of information has an obvious limitation. The algorithm for dynamically segmenting system presentations based on information measure has no knowledge of the content of the presentation and the discourse structure of the presentation. Therefore, semantically related information may be presented into separated segments, and semantically different information may be grouped together into one segment. One future direction of the current work is to incorporate information control with discourse structure, so that system responses will be presented coherently and with the right amount of information.

Though the user study shows that the usability of the system can be improved by simply planning how much information to present during system presentations, it also demonstrates a need for interaction. First, even when the user has unlimited access to the system, sometimes the user still can not get perfect recall and precision. Most problems are with names, which are hard to recognize. In such cases, the user prefers to interrupt the system, and specifically asks for clarifications. Second, since not all information is of interest to the user, the user has to listen through much irrelevant information in order to get the relevant information. Interaction with the user will allow the system to incrementally develop user models and tailor descriptions to the user's interest. Third, users prefer immediate feedback. If the users are not clear about certain facts, they prefer to ask clarification questions right after the presentation, rather than waiting for complete playbacks.

Summary and Future Work

Where a system should place the turn transition relevance points is an important design issue in human-computer interaction. In this paper, the amount of information in presentations is investigated as one factor for planning TRPs. A preliminary user study indicates that the placement of TRPs can be optimized based on the amount of information in presentations.

TRP Setting	System Run	Time Per 100 Words	Precision	Recall
5	1	67.3	100%	100%
10	1	66.7	100%	100%
20	1	53.8	100%	73%
	2	107.8	100%	91%
40	1	47.3	83%	70%
	2	94.6	88%	88%
	3	141.9	88%	88%
50	1	44.9	96%	68%
	2	85.3	100%	78%
	3	134.6	100%	100%

Table 2: User Study Results

We intend to conduct more user studies to verify the preliminary results. An obvious limitation of the current approach is that the information measure has no knowledge of the content of the presentation and the discourse structure of the presentation. One future direction of our work is to incorporate information control with discourse structure, so that system responses will be presented coherently and with the right amount of information. Another direction of our work is to study information control in interactive settings and its implications for content planning and interaction management.

Acknowledgements

Sincere thanks to the volunteers for participating in the user study. I thank Lori Levin and Jaime Carbonell for discussion about the ideas in the paper.

References

Abella, A.; Brown, M. K.; and Buntschuh, B. 1996. Development principles for dialogue-based interface. In *Proceedings of ECAI'96 Workshop on Spoken Dialogue Systems*, 1–7.

Brennan, S. E.; Friedman, M. W.; and Pollard, C. 1987. A centering approach to pronouns. In *Proceedings of the 25th ACL.*

Cawsey, A. 1992. *Explanation and Interaction.* Cambridge, Massachusetts: The MIT Press.

Grishman, R., and Sundheim, B. 1996. Message understanding conference - 6: A brief history. In *Proceedings of COLING 1996.*

Grosz, B. J., and Sidner, C. J. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics* 12 (3).

Guindon, R.; Sladky, P.; Brunner, H.; and Conner, J. 1986. The structure of user-adviser dialogues: Is there method in their madness? In *Proceedings of 24th ACL.*

Haller, S. 1994. Recognizing digressive questions. In *Proceedings of AAAI94 Fall Symposium.*

Haller, S. 1995. Planning text interactively using plan recognition. In *Proceedings of IJCAI workshop on plan recognition.*

Hirschberg, J., and Pierrehumbert, J. B. 1986. The intonational structure of discourse. In *Proceedings of the 24th ACL.*

Maybury, M. T. 1991. *Planning Multisentential English Text Using Communicative Acts.* Ph.D. Dissertation, Computer Laboratory, University of Cambridge.

McKeown, K. R. 1985. *Text Generation.* Cambridge, UK: Cambridge University Press.

Moore, J. D., and Paris, C. 1992. Planning text for advisory dialogues: capturing intentional and rhetorical information. *Computational Linguistics* 19 (4):651–695.

Reichman, R. 1985. *Getting Computers to Talk like You and Me.* Cambridge, Mass.: MIT Press.

Sacks, H.; Schegloff, E. A.; and Jefferson, G. 1978. A simplest systematics for the organization of turn-taking for conversation. *Language* 50:696–735.

Traum, D. R., and Heeman, P. A. 1996. Utterance units in spoken dialogue. In *ECAI Workshop on Dialogue Processing in Spoken Language Systems.*

Whittaker, S., and Stenton, P. 1988. Cues and control in expert-client dialogues. In *Proceedings of 26th ACL*, 123–130.