ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Coherent Topic Transition in a Conversational Agent

Daniel Macias-Galindo, Wilson Wong, Lawrence Cavedon, John Thangarajah

School of Computer Science and Information Technology
RMIT University, Melbourne, Australia

{daniel.macias,wilson.wong,lawrence.cavedon,john.thangarajah}@rmit.edu.au

10.21437/Interspeech.2012-230

Abstract

A conversational agent for entertainment and engagement requires the ability to maintain coherent conversations. We describe the use of *semantic relatedness* to select the next conversational fragment that an agent utters, to maximise dialogue coherence and/or to suggest new directions for a dialogue. We compare our approach, using a specific semantic relatedness metric, to an existing *nearest-context* mechanism based on $TF \times IDF$ for selecting fragments to continue a conversation. Evaluation with human judges shows that use of semantic relatedness provides improved coherence across a sample collection of generated conversations.

Index Terms: Spoken dialogue, dialogue coherence, semantic relatedness

1. Introduction

The majority of the recent work on spoken language dialogue systems has been in the context of *task-oriented dialogue*, for instance TRIPS [1, 2]. Unlike *chatbots* which sit at the other end of the spectrum, the range of utterances that task-oriented systems are expected to recognise and respond to are structured by the associated tasks. The work in this paper is set in the context of a conversational agent embodied as a child's Toy that mixes tasks (e.g. telling stories, playing games) with "chatter" that is designed to entertain and (ultimately) build engagement and relationship. The task-oriented aspect of the conversational Toy is implemented as *plans* in the *BDI (Belief-Desire-Intention)* framework of intelligent agents [3]. We focus here on the chat capabilities of the Toy, and in particular, the challenge of maintaining a coherent conversation *with the agent taking initiative*.

As opposed to the standard chatbot approach (e.g. ALICE [4]) which is designed to respond to any user utterance using generic responses and deflection strategies, our conversational Toy takes initiative for chat in order to structure the conversation and increase the likelihood of understanding input [3]. Our language processing framework is relatively simple and relies on *conversational fragments*, which are short snippets of conversation associated with a topic. Each fragment predicts a range of inputs and how to interpret them. This is not dissimilar to a simple VoiceXML (VXML) approach to interpreting input. However, unlike a typical VXML application, there is no task to structure a dialogue around. As such, the authoring of long fragments, and the prediction of inputs and where the dialogue goes next, become infeasible.

In this paper, we present and evaluate the use of *semantic relatedness* by our conversational agent for joining short snippets into longer conversations while maximising coherence. We use semantic relatedness, which is a standard component of many measures of *discourse coherence* [5], as part of the process of selecting the next conversational fragment, when required, as opposed to measuring the coherence of a whole text fragment.[1] Our method is related to the approach used by Gandhe and Traum [7] to select the next system utterance from amongst a collection of authored candidates. However, their mechanism is based on a different metric (i.e. $TF \times IDF$) and is only used to select an appropriate utterance, whereas our aim is to select an appropriate new direction for a dialogue.

2. The Conversational Toy

In this section we briefly outline the architecture of the conversational Toy [3] that sets the context for this work. The full system involves automatic speech recognition (ASR) and text-to-speech (TTS) synthesis for receiving inputs and generating responses. The Toy contains a *Dialog Manager* (DM) for managing the interaction between the Toy and the user. It comprises three components, namely, *Input Handling* (IH), *Conversation Management* (CM) and *Activity Management* (AM). The IH component analyses and extracts weighted key phrases, topics and sentiments from the user inputs, while the AM manages the selection, focus and interaction between *Conversational Activities* such as story-telling and quizzes. The CM manages the interpretation of input and utterance selection from the conversational activities. Each conversational activity is developed as a module to guide conversations around an activity. The modules encapsulate: domain knowledge; conversational fragments; rules to man-

[1] Measuring local coherence (i.e. utterance to utterance) ultimately leads to global coherence over a dialogue [6].

age the activities of the module; and an *input grammar* which specifies the trigger (input from the user) for entering an activity. In this work, we describe and evaluate the part of the CM that is responsible for selecting the next conversational fragment (i.e. system utterance) whilst maintaining coherence with respect to the topic.

The fragments are developed by *activity module* designers. Each fragment comprises two aspects: (a) a *header* containing a unique identifier and a set of string *key terms* used for classification; and (b) a *body* that contains the output text of the fragment (or a template for such) plus a list of *expected response* templates to match possible user responses to the output. The expected responses are used by the DM to match against user inputs to determine the progression of conversations, and may have an associated pointer to a successor fragment. In the experiment detailed in this paper, we use a set of multi-topic conversational fragments authored by creative writing students.[2] Such fragments have been constructed to form a story line; that is, transitions to other fragments are pre-scripted and activated by terms from the user inputs, while the terms in the fragments used in the dialogue contribute to the *context*, which is described in the next section. Fragments for this experiment are classified into three types: conversation *starters*, *finalisers* and *continuers*.

A *starter* fragment s is one which no another fragment points to, while a *finaliser* fragment f does not point to other fragments. A *continuer* fragment c contains both properties, while dynamically contributing to the enrichment of the dialogue context due to specific terms contained in the previous user input. The set of all starters is denoted as $S = \{s_1, ..., s_n\}$. In principle, each starter fragment s links to different continuers depending on the matches between the user inputs and the expected inputs specified in the body of s. Depending on the next fragments associated with the expected inputs, conversations will progress differently. Any continuation c can also link to multiple other continuers or finalisers through this setup, forming a network of interconnected fragments.

3. Topic Selection using Semantic Relatedness

Existing approaches for measuring coherence are based on the analysis of *lexical* and *semantic* features [5, 7, 6]. In this section, we describe our approach to selecting the next utterance given a completed conversation snippet. Our approach is governed by the principle of *coherence* as a property of linguistic acts. For instance, in Figure 1, the upper dialogue can be considered coherent as the new utterance flows around a main context, while the lower dialogue is labelled incoherent due to the sudden jump to unrelated topics.

U1: What is your favourite sport?
U2: Soccer
U1: You must need a lot of energy to play that sport. Where do you get all your energy?
U2: I eat a lot
U1: That's good.
$< topic_suggestion >$
U1: What's your favourite energy food?

(a)

U1: When you go to a restaurant with mummy and daddy, what do you order?
U2: Hot chips and sauce
U1: You're making me very hungry!
$< topic_suggestion >$
U1: How about creating a superhero?

(b)

Figure 1: Examples of (a) coherent; and (b) less coherent topic selections for dialogue acts.

Given the setting of our conversational agent, where outputs are selected from a set of predefined starter fragments, the scenario shown in Figure 1(b) is equally likely to result unless output utterances are chosen in a way that accounts for coherence. We use semantic relatedness to conduct the selection process as follows. First, each fragment has a set of tag words (i.e. nouns) associated with it, denoted as T. The tag words from the most recent two utterances are used as context (following [7]), denoted as κ. Second, once a finaliser is reached, a set of all other candidate starters to choose from, denoted as S, is compiled. According to our semantic relatedness approach, the most coherent starter $s \in S$ for maintaining the dialogue is defined as the one that has the highest *group-average* relatedness $Rel(\kappa, T)$ between its set of tag words T and the context set κ. The group-average relatedness is the average over the pairwise relatedness of terms taken from two sets. We use Equation 1 to measure the group-average relatedness between the tag words of the candidate starter $s_r \in S$ and the context κ:

$$Rel(\kappa, T_r) = \frac{\sum_{\forall k \in \kappa, \forall t \in T_r} NWR(k,t)}{|\kappa||T_r|} \quad (1)$$

where $NWR(k,t) = e^{-0.6 \times NWD(k,t)}$, which stands for *Normalised Web Relatedness*. *Normalised Web Distance (NWD)* [8] is defined in Equation 2, where G is the number of English articles in Wikipedia. We have previously [9] found this distributional semantic metric to provide robustness and broad coverage.

$$NWD(k,t) = \frac{\log(\max(|k|,|t|) - \log(|k \cap t|)}{\log(|G|) - \log\min(|k|,|t|)} \quad (2)$$

Note that much of the information required for this calculation can be precomputed since the fragments and

[2] We have recently developed techniques to automatically mine conversational fragements from web forums.

their terms are known in advance. Relatedness with terms in the current context can be computed while responding and waiting for next user input.

4. Evaluation

We evaluate the use of semantic relatedness ($semrel$) for selecting a starter conversational fragment after a finaliser is reached, by comparing it against two alternatives: (1) making an uninformed or random selection ($random$) amongst candidate starter fragments, and (2) making a selection using Gandhe and Traum's nearest-context ($nearct$) mechanism [7]. The three approaches are compared using human judgements of the level of "coherence" of transitioning from the preceding finalisers, over a given collection of generated dialogues. Using the $random$ mechanism, a fragment is selected randomly from a set of available starters. For the $nearct$ approach, $TF \times IDF$ is used to select a fragment that is the most *similar* to the preceding context [7]. Our $semrel$ mechanism uses NWR over Wikipedia, as described above, to select the fragment that contains terms that are more semantically related to the current context, based on the semantic view of text coherence proposed by Lapata and Barzilay [5].

Experimental setup: For our evaluation, only fragments that form linear conversations are used. That is, expected responses of such a fragment all link to the same next fragment. We identified a set of 13 starters, S, from our collection of fragments that conform to this condition. For each starter $s_x \in S$, a succession of continuers would follow, ending with a finaliser f_x to form a sequence $s_x \rightarrow c_1 \rightarrow c_2 \rightarrow ... \rightarrow c_m \rightarrow f_x$. From this set S of 13 starters, we constructed 12 distinct conversations, which are essentially alternations between the output utterances (i.e. system utterance) and randomly selected expected inputs (i.e. simulated user input) from every fragment in the respective sequences. One of the starters was removed from the collection because it produces a very similar dialogue sequence to another starter, where only the gender of the child was different (i.e. *boy* vs *girl*). This starter was used in the process for continuing a conversation, however.

The three approaches discussed above were used to extend these 12 distinct conversations, as follows. For a conversation that starts with s_x, there are $|S| - 1$ other starters from the set $S - \{s_x\}$ to choose from to determine the most coherent re-starter from f_x. For that particular conversation, the last two system utterances f_x and c_m are used to produce the context for that conversation. For the $semrel$ approach, the context is constructed from terms found in the fragment head of such utterances. As for the $nearct$ approach, the context comprises all the terms available in the previous two utterances (as per [7]). These mechanisms then select a fragment $s_y \in S - \{s_x\}$ as the most coherent starter to re-start from where the previous sequence (starting with s_x) ended.

We presented the system utterances and the simulated user inputs from c_m and f_x, along with the three alternative transitions generated using the three different approaches $random$, $nearct$ and $semrel$ for all 12 conversations to human judges, to be rated for "coherence" of the topic transition.[3] We obtained judgements from ten judges assessing the coherence of the part of the conversation represented by $c_m \rightarrow f_x \rightarrow s_y$ for the different generated alternative transitions. The judges used a 5-point Likert scale, with 0 representing `highly incoherent` and 4 representing `highly coherent`. A description justifying each score was also requested. Each judge assessed coherence for a total of 32 interactions (i.e. 12 conversations $\times$ 3 potential transitions generated using three approaches, with 4 cases where $semrel$ and $nearct$ produced the same output).

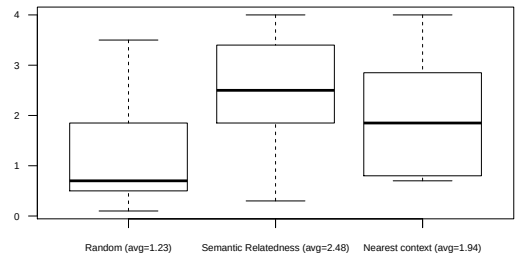


Figure 2: Distribution of scores for each approach.

Results and discussion: We collected and analysed the assessments; the results are shown in Figure 2 as boxplots for each of the fragment-selection mechanism used. Space precludes showing individual scores for each judge, but the averages for each mechanism across all judges are shown in the figure. Average Pearson inter-assessor correlation, i.e. the Pearson correlation between each judge and the average of the other judges, was calculated as $\rho = 0.82$, indicating strong agreement across the ten human judges.

Figure 2 summarises the average rating achieved by the transitions using the respective approaches for all 12 conversations. Unsurprisingly, the $random$ mechanism produced the least coherent transitions, as deemed by the human assessors, with the median sitting below the scale of 1 (i.e. `incoherent`). This, together with the sample maximum (i.e. the upper arm or whisker of the box) reaching above scale 3, shows the erratic nature of the $random$ approach, which is as expected. For the $nearct$ approach, the median lies on the scale of 2, which is `neither incoherent nor coherent`. Considering the symmetrical nature of the lower and upper

[3]There were a small number of cases where the starters chosen using $semrel$ and $nearct$ were similar; in this situations, only two approaches were displayed in the survey.

quartile of the box, the coherence of the continuers determined using this approach cannot be guaranteed. Our $semrel$ approach, on the other hand, has the lower quartile on and above the scale of 2. In other words, the majority of the transitions using the $semrel$ approach (75%) were rated as better than neutral by the judges. To determine if the difference between the three techniques is statistically different, we performed the Wilkoxon rank sum test for all the user judgements. It was found that the three techniques are significantly different (with confidence value of 0.99), where $random$ performs the poorest with the most significant difference, while the difference between $nearct$ and $semrel$ is also significant at the 0.99 level. This difference is observable in Figure 2.

Figure 3(a) shows a specific conversation from our experiment where $semrel$ outperforms $nearct$ by maintaining the topic of the conversation (i.e. food) instead of an abrupt change with a starter about activities.

SYS: Do you ever get snacks from the canteen?
USR: Yes
SYS: What snack do you like to buy?
USR: Mixed Lollies
SYS: ok, let's talk about something else
$< candidate_transitions >$
(nearct) SYS: Do you like activities?
(semrel) SYS: What is your favourite cereal?

(a)

SYS: You know what I think would be **beauti**fully **magic**al? To see a wonderful **mermaid** at the **beach**.
USR: Yes that would be terrific.
SYS: I'd definitely agree with you.
$< candidate_transitions >$
i) SYS: What do you like best about being a **girl**?
ii) SYS: Do you believe in **magic** or make-believe?

(b)

Figure 3: (a) A conversation where the nearest context approach results in lower coherence compared to the selection made by our semantic relatedness mechanism; and (b) an example where higher significance of terms is not recognised; significant terms are shown in bold.

Error analysis: Three dialogues using the $semrel$ mechanism fell below a score of 2 of the scale. In some cases, the fact of having multiple terms per utterance diminished the effect of the most significant keywords – Figure 3(b) illustrates this situation. In this case, there are two candidate transitions (i.e. follow-on utterances) after the conversational thread ends. Judges considered (ii) a more appropriate selection, due to a link between magic and mermaids; however, terms such as *beauty-girl* and *mermaid-girl* were deemed to have higher relatedness than *beauty-magic* and *mermaid-magic*. Thus, the semantic relatedness approach selected phrase (i). This example also illustrates another problem, which is an issue with our shallow processing of conversational fragments, where the statement about "being a girl" may be selected regardless of the child or even previously stated information about him/her. [7] describes a *segmented nearest-context* approach, which activates specific utterances when certain terms are mentioned. Augmenting our selection mechanism with something similar, especially when salient information becomes known, could be used to address this sort of problem.

5. Conclusions

We have described an approach, based on a semantic relatedness mechanism for topic-coherence, to transition between conversational snippets in an agent that engages in "chatty" dialogue. Our evaluation suggests that this mechanism results in more coherent topic switches than the mechanism for selecting output utterances presented in [7]. In more recent work, we have developed techniques for automatically mining conversational content from Web forums and assembling these disjointed fragments into coherent conversation using an approach dependent on the techniques we have described here. As immediate Future Work, we intend to improve on the identification of the most significant topic terms and semantic relationships to address examples such as those in Figure 3(b).

Acknowledgements: This work is partially supported by the Australian Research Council and Real Thing Entertainment Pty. Ltd. under grant number LP110100050. Daniel Macias-Galindo is sponsored by CONACYT, Mexico, scholarship number 201228.

6. References

[1] J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, "Toward conversational human-computer interaction", *AI Magazine*, vol. 22, no. 4, 2001.

[2] M. O. Dzikovska, J. F. Allen, and M. D. Swift, "Linking semantic and knowledge representations in a multi-domain dialogue system", *Journal of Logic and Computation*, vol. 18, 2008.

[3] W. Wong, L. Cavedon, J. Thangarajah, and L. Padgham, "Flexible conversation management using a BDI agent approach", in *IVA*, Santa Cruz CA, 2012.

[4] R. S. Wallace, "The anatomy of A.L.I.C.E.", in *Parsing the Turing Test*, R. Epstein, G. Roberts, and G. Beber, Eds. Springer, 2009.

[5] M. Lapata and R. Barzilay, "Automatic Evaluation of Text Coherence: Models and Representations", in *IJCAI*, vol. 19, 2005.

[6] A. Purandare and D. Litman, "Analyzing dialog coherence using transition patterns in lexical and semantic features", in *International FLAIRS Conference*, 2008.

[7] S. Gandhe and D. Traum, "Creating spoken dialogue characters from corpora without annotations", in *INTERSPEECH*, 2007.

[8] R. L. Cilibrasi and P. M. Vitanyi, "The Google Similarity Distance", *Trans. Knowledge and Data Engineering*, vol. 19, 2007.

[9] W. Wong, W. Liu, and M. Bennamoun, "Featureless data clustering,", in *Handbook of Research on Text and Web Mining Technologies*, M. Song and Y. Wu, Eds. IGI Global, 2008.