Telling questions from statements in spoken dialogue systems

Sofia Strömbergsson, David House, Jens Edlund
KTH Speech, Musik and Hearing
Stockholm, Sweden
sostr@csc.kth.se, (edlund,davidh)@speech.kth.se

1. Domain dependency: controlling context, style and situation

To date, just about every system that successfully communicates with humans using language owes a great deal of its success to one particular set of characteristics. The ability to control or make assumptions about the linguistic context, the dialogue style and the physical situation in which the dialogue occurs is essential to systems ranging from early text based systems such as Eliza (Weizenbaum, 1966) to modern speech based question-answer systems such as Siri, acquired by Apple in 2010 and launched the companies’ smart phones in 2011. Controlling or predicting context, style and situation allows us to build domain dependent systems – systems designed to handle a very small subset of the dialogues humans in which take part.

One of the motivations for spoken dialogue system research is an improved understanding of human interaction. But demonstrating an understanding of human interaction by mimicking it in a machine requires that we do not constrain the domain any more than humans do. A widely held hope is that domain independency can be reached by (1) gradually increasing in-domain coverage, (2) gradually widening the domains, (3) adding new domains, (4) developing robust domain detection, and (5) methods to seamlessly move from one domain to another. The soothing idea is that progress can be gradual, and that each small improvement adds to the improvement of the whole.

When subjected to closer scrutiny, the hope of gradually building ourselves out of domain dependency may be overly optimistic. Some distinctions that seem trivial to humans and spoken dialogue systems alike may turn out to be unexpectedly difficult when we relinquish control over context, style and situation. We suspect that the simple distinction between question and statement is an example of this.

2. Dialogue acts: questions and statements

Deciding whether an utterance is a question or a statement is relatively uncontroversial for humans. In some cases, questions are clearly marked lexically, syntactically, or both: wh-questions have an initial wh-word, and yes/no questions have their tell-tale word order. In other cases, utterances that are pragmatic questions – utterances that function as a question – lack these distinctive features and utterances that are superficially questions do not have question function. Yet people rarely find their understanding problematic. The following dialogue snippet from a fictitious interview in the 2000 film How to kill your neighbor’s dog is an exception:

Debra: There’s no denying that in the 80s you were quite the boy wonder. All your plays went on to Broadway. That hasn’t been the case in the 90s, you’ve had something like three bombs in a row. [SILENCE]. Peter?
Peter: I didn’t hear a question in there.

Statements like Peter’s are common in political debate, citations from which make up a substantial part of the 100000+ hits Google delivers for “I did not hear/detect a question”. Outside of rhetoric and comedy, however, it is difficult to find examples of people mistaking questions for statements and vice versa.

3. Questions and spoken dialogue systems

As with people, telling questions and statements apart is not on any list of hard problems for current spoken dialogue systems. This, however, is largely because the systems are designed so that speech acts are predictable. There is a correlation between initiative and question, so that a dialogues with system initiative will rarely allow questions from the user, so user questions will not be understood and the non-understanding will be explained as out-of-domain. In the case of user-initiative dialogue, the opposite is true, and most user utterances are expected to be questions. Even mixed-initiative dialogues often constrain the users' freedom. In the mixed-initiative apartment browsing system Adapt (Gustafson et al., 2000), for example, a dialogue consisted of two phases. In the first phase, the system would ask the user to provide constraints to narrow down the search. When the number of suitable apartments was 7 or less, the system entered the second phase, in which the user asks questions about the apartments, and the system answers.

These tricks are lost in a move towards humanlike systems engaging in everyday conversations with people, and the question/statement distinction becomes less trivial for spoken dialogue systems. For this reason, we investigate 600 questions and question-like utterances from the Spontal corpus of conversational Swedish with a view to cues as to how a spoken dialogue system can best tell them apart from statements.

4. Hypothesis

This work is preliminary, but it takes as its starting point the analyses we have already done on questions in the Spontal corpus, which suggest in particular that (1) the concept of question is difficult to define and (2) question markers are not always present. Dialogue act classifications are often quite circular when it comes to defining questions. As an example, speech act classification used in the Edinburgh Map Task (Anderson et al., 1991), for uses “question” both in label and description and hedges with a catch-all definition of wh-questions: “Any query not covered by the other categories”. We expect to find, then, that the identification of questions is far from easy, but rather is one of the key problems to solve for spoken dialogue systems to take a leap towards achieving human conversational skills.
5. Material

Questions and matching non-questions in this study were extracted from orthographically transcribed subset of 24 dialogues from the Spontal corpus (Edlund et al., 2010). Two annotators looked for questions while transcribing the spoken dialogues and labeled these with a question tag. The definition of “question” was deliberately kept quite open: “Anything that resembles structurally or functionally, in whole or in part, a question”. In all, 908 talkspurts received the question label.

Three independent annotators labeled all 908 instances with respect to four relatively simple queries. (For a detailed description of the queries, see Strömbergsson et al., in press-a). During this process, annotators could choose to skip talkspurts that they felt were in no sense a question, or that were otherwise impossible to judge. Talkspurts that were skipped by at least one annotator, or were all annotators that disagreed on the question label, were excluded from further analysis, leaving a set of 641 questions. The targeted 600 questions were selected from this set so that they were balanced for the interlocutors’ gender and previous acquaintance.

For each of the 600 questions, a matching non-question was selected from the same speaker and the same dialogue as the question, with the restriction that a) the difference in duration between the question and the non-question, and b) the time between the question and the non-question in the dialogue, were minimized. (Talkspurts that had been labeled with a question tag earlier were not eligible for selection.)

6. Prosodic measures

Five different prosodic measures were extracted from the questions and non-questions: speech rate (RATE), average pitch (AVG), pitch variation (VAR), pitch range (RNG), and an estimate of intonation slope (DIFF). (For descriptions of these measures, see Strömbergsson et al., in press-b).

7. Results

In order to explore prosodic (within-speaker) differences between questions and matching non-questions, multiple t-tests were conducted. Table 1 shows the results of these tests.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Qns</th>
<th>nQns</th>
<th>t-test statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATE</td>
<td>M</td>
<td>SD</td>
<td>t(599) = 12.0, p &lt; .001</td>
</tr>
<tr>
<td>AVG</td>
<td>M</td>
<td>SD</td>
<td>t(599) = 3.7, p &lt; .001</td>
</tr>
<tr>
<td>VAR</td>
<td>M</td>
<td>SD</td>
<td>t(599) = -.8, p = .42</td>
</tr>
<tr>
<td>RNG</td>
<td>M</td>
<td>SD</td>
<td>t(599) = -1.0, p = .32</td>
</tr>
<tr>
<td>DIFF</td>
<td>M</td>
<td>SD</td>
<td>t(599) = -.5, p = .60</td>
</tr>
</tbody>
</table>

Table 1: Descriptive and t-test statistics for the five prosodic measures, for questions and non-questions.

The figures presented in Table 1 suggest that questions are produced faster than non-questions, and with slightly higher pitch. No other differences are found.

The finding that questions are produced at a faster speaking rate than non-questions is surprising and could be explained by the fact that the questions have all been selected so that they contain nothing more than the actual question, whereas we do not have the same control for the non-questions. The non-questions can (and do) also contain other vocalizations, such as laughter and breathing, which contributes to longer duration but without an accompanying increase in the number of consonant-to-vowel transitions. Thus, speaking rate appears to be slower.

8. Discussion and future work

The two-page abstract format for SLTC publications does not provide room for further detail, nor for discussion. We are forced refer the reader to the accompanying poster presentation and to forthcoming full publications.

9. Acknowledgements

This work is funded by the Swedish Research Council, Humanities and Social Sciences (VR 2009-1764) “Intonational variation in questions in Swedish”.

10. Reference


