Designing a Task-Based Evaluation Methodology for a Spoken Machine Translation System

Kavita Thomas
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
kavita@cs.cmu.edu

Abstract

In this paper, I discuss issues pertinent to the design of a task-based evaluation methodology for a spoken machine translation (MT) system processing human to human communication rather than human to machine communication. I claim that system mediated human to human communication requires new evaluation criteria and metrics based on goal complexity and the speaker's prioritization of goals.

1 Introduction

Task-based evaluations for spoken language systems focus on evaluating whether the speaker's task is achieved, rather than evaluating utterance translation accuracy or other aspects of system performance. Our MT project focuses on the travel reservation domain and facilitates on-line translation of speech between clients and travel agents arranging travel plans. Our prior evaluations (Gates et al., 1996) have focused on end-to-end translation accuracy at the utterance level (i.e., fraction of utterances translated perfectly, acceptably, and unacceptably). While this method of evaluation conveys translation accuracy, it does not give any information about how many of the client's travel arrangement goals have been conveyed, nor does it take into account the complexity of the speaker's goals and task, or the priority that they assign to their goals; for example, the same end-to-end score for two dialogues may hide the fact that in one dialogue the speakers were able to communicate their most important goals while in the other they were only able to communicate successfully the less important goals.

One common approach to evaluating spoken language systems focusing on human-machine dialogue is to compare system responses to correct reference answers; however, as discussed by (Walker et al., 1997), the set of reference answers for any particular user query is tied to the system's dialogue strategy. Evaluation methods independent of dialogue strategy have focused on measuring the extent to which systems for interactive problem solving aid users via log-file evaluations (Polifroni et al., 1992), quantifying repair attempts via turn correction ratio, tracking user detection and correction of system errors (Hirschman and Pao, 1993), and considering transaction success (Shriberg et al., 1992). (Danieli and Gerbino, 1995) measure the dialogue module's ability to recover from partial failures of recognition or understanding (i.e., implicit recovery) and inappropriate utterance ratio; (Simpson and Fraser, 1993) discuss applying turn correction ratio, transaction success, and contextual appropriateness to dialogue evaluations, and (Hirschman et al., 1990) discuss using task completion time as a black box evaluation metric.

Current literature on task-based evaluation methodologies for spoken language systems primarily focuses on human-computer interactions rather than system-mediated human-human interactions. For a multilingual MT system, speakers communicate via the system, which translates their responses and generates the output in the target language via speech synthesis. Measuring solution quality (Sikorski and Allen, 1995), transaction success, or contextual appropriateness is meaningless, since we are not interested in measuring how efficient travel agents are in responding to clients' queries, but rather, how well the system conveys the speakers' goals. Likewise, task completion time will not capture task success for MT dialogues since it is dependent on dialogue strategies and speaker styles. Task-based evaluation methodologies for

MT systems must focus on whether goals are communicated, rather than whether they are achieved.

2 Goals of a Task-Based Evaluation Methodology for an MT System

The goal of a task-based evaluation for an MT system is to convey whether speakers' goals were translated correctly. An advantage of focusing on goal translation is that it allows us to compare dialogues where the speakers employ different dialogue strategies. In our project, we focus on three issues in goal communication: (1) distinction of goals based on subgoal complexity, (2) distinction of goals based on the speaker's prioritization, and (3) distinction of goals based on domain.

3 Prioritization of Goals

While we want to evaluate whether speakers' important goals are translated correctly, this is sometimes difficult to ascertain, since not only must the speaker's goals be concisely describable and circumscribable, but also they must not change while she is attempting to achieve her task. Speakers usually have a prioritization of goals that cannot be predicted in advance and which differs between speakers; for example, if one client wants to book a trip to Tokyo, it may be imperative for him to book the flight tickets at the least, while reserving rooms in a hotel might be of secondary importance, and finding out about sights in Tokyo might be of lowest priority. However, his goals could be prioritized in the opposite order, or could change if he finds one goal too difficult to communicate and abandons it in frustration.

If we insist on eschewing unreliability issues inherent in asking the client about the priority of his goals after the dialogue has terminated (and he has perhaps forgotten his earlier priority assignment), we cannot rely on an invariant prioritization of goals across speakers or across a dialogue. The only way we can predict the speaker's goals at the time he is trying to communicate them is in cases where his goals are not communicated and he attempts to repair them. We can distinguish between cases in which goal communication succeeds or fails, and we can count the number of repair attempts in both cases. The insight is that speakers will attempt to repair higher priority goals more than lower priority goals, which they will abandon sooner. The number of repair attempts per goal quantifies the speaker's priority per goal to some degree.

We can capture this information in a simple metric that distinguishes between goals that eventually succeed or fail with at least one repair attempt. Goals that eventually succeed with t_g repair attempts can be given a score of $1/t_g$, which has a maximum score of 1 when there is only one repair attempt, and decays to 0 as the number of repair attempts goes to infinity. Similarly, we can give a score of $-(1 - 1/t_g)$ to goals that are eventually abandoned with t_g repair attempts; this has a maximum of 0 when there is only a single repair attempt and goes to -1 as t_g goes to infinity. So the overall dialogue score becomes the average over all goals of the difference between these two metrics, with a maximum score of 1 and a minimum score of -1.

$$score(goal) = \begin{cases} \frac{1}{t_g} & \text{for successful goal} \\ -(1 - \frac{1}{t_g}) & \text{for unsuccessful goal} \end{cases} \quad (1)$$

$$score(dialogue) = \frac{1}{numgoals} \sum_{goals} score(goal) \quad (2)$$

4 Complexity of Goals

Another factor to be considered is goal complexity; clearly we want to distinguish between dialogues with the same main goals but in which some have many subgoals while others have few subgoals with little elaboration. For instance, one traveller going to Tokyo may be satisfied with simply specifying his departure and arrival times for the outgoing and return laps of his flight, while another may have the additional subgoals of wanting a two-day stopover in London, vegetarian meals, and aisle seating in the non-smoking section. In the metric above, both goals and subgoals are treated in the same way (i.e., the sum over goals includes subgoals), and we are not weighting their scores any differently.

While many subgoals require that the main goal they fall under be communicated for them to be communicated, it is also true that for some speakers, communicating just the main goal and not the subgoal may be a communication failure. For example, if it is crucial for a speaker

to get a stopover in London, even if his main goal (requesting a return flight from New York to Tokyo) is successfully communicated, he will view the communication attempt a failure unless the system communicates the stopover successfully also. On the other hand, communicating the subgoal (e.g., a stopover in London), without communicating the main goal is nonsensical – the travel agent will not know what to make of "a stopover in London" without the accompanying main goal requesting the flight to Tokyo.

However, even if two dialogues have the same goals and subgoals, the complexity of the translation task may differ; for example, if in one dialogue (A) the speaker communicates a single goal or subgoal per speaker turn, while in the other (B) the speaker communicates the goal and all its subgoals in the same speaker turn, it is clear that the dialogue in which the entire goal structure is conveyed in the same speaker turn will be the more difficult translation task. We need to be able to account for the average goal complexity per speaker turn in a dialogue and scale the above metric accordingly; if dialogues A and B have the same score according to the given metric, we should boost the score of B to reflect that it has required a more rigorous translation effort. A first attempt would be to simply multiply the score of the dialogue by the average subgoal complexity per main goal per speaker turn in the dialogue, where N_{mg} is the number of main goals in a speaker turn and N_{sg} is the number of subgoals. In the metric below, the average subgoal complexity is 1 for speaker turns in which there are no subgoals, and increases as the number of subgoals in the speaker turn increases.

$$score'(dialogue) = score(dialogue) * \frac{1}{numspkturns} \sum_{spkturns} [\frac{N_{sg} + N_{mg}}{N_{mg}}] \quad (3)$$

5 Our Task-Based Evaluation Methodology

Scoring a dialogue is a coding task; scorers will need to be able to distinguish goals and subgoals in the domain. We want to minimize training for scorers while maximizing agreement between them. To do so, we list a predefined set of main goals (e.g., making flight arrangements or hotel bookings) and group together all subgoals that pertain to these main goals in a two-level tree. Although this formalization sacrifices subgoal complexity, we are unable to determine this without predefining a subgoal hierarchy and we want to avoid predefining subgoal priority, which is set by assigning a subgoal hierarchy.

After familiarizing themselves with the set of main goals and their accompanying subgoals, scorers code a dialogue by distinguishing in a speaker turn between the main goals and subgoals, whether they are successfully communicated or not, and the number of repair attempts in successive speaker turns. Scorers must also indicate which domain each goal falls under; we distinguish goals as in-domain (i.e., referring to the travel-reservation domain), out-of-domain (i.e., unrelated to the task in any way), and cross-domain (i.e., discussing the weather, common polite phrases, accepting, negating, opening or closing the dialogue, or asking for repeats).

The distinction between domains is important in that we can separate in-domain goals from cross-domain goals; cross-domain goals often serve a meta-level purpose in the dialogue. We can thus evaluate performance over all goals while maintaining a clear performance measure for in-domain goals. Scores should be calculated separately based on domain, since this will indicate system performance more specifically, and provide a useful metric for grammar developers to compare subsequent and current domain scores for dialogues from a given scenario.

In a large scale evaluation, multiple pairs of speakers will be given the same scenario (i.e., a specific task to try and accomplish; e.g., flying to Frankfurt, arranging a stay there for 2 nights, sightseeing to the museums, then flying on to Tokyo); domain scores will then be calculated and averaged over all speakers.

Actual evaluation is performed on transcripts of dialogues labelled with information from system logs; this enables us to see the original utterance (human transcription) and evaluate the correctness of the target output. If we wish to, log-file evaluations also permit us to evaluate the system in a glass-box approach, evaluating individual system components separately (Simpson and Fraser, 1993).

6 Conclusions and Future Work

This work describes an initial attempt to account for some of the significant issues in a task-based evaluation methodology for an MT system. Our choice of metric reflects separate domain scores, factors in subgoal complexity and normalizes all counts to allow for comparison among dialogues that differ in dialogue strategy, subgoal complexity, number of goals and speaker-prioritization of goals. The proposed metric is a first attempt, and describes work in progress; we have attempted to present the simplest possible metric as an initial approach.

There are many issues that need to be addressed; for instance, we do not take into account optimality of translations. Although we are interested in goal communication and not utterance translation quality, the disadvantage to the current approach is that our optimality measure is binary, and does not give any information about how well-phrased the translated text is. More significantly, we have not resolved whether to use metric (1) for both subgoals and goals together, or to score them separately. The proposed metric does not reflect that communicating main goals may be essential to communicating their subgoals. It also does not account for the possible complexity introduced by multiple main goals per speaker turn. We also do not account for the possibility that in an unsuccessful dialogue, a speaker may become more frustrated as the dialogue proceeds, and her relative goal priorities may no longer be reflected in the number of repair attempts. We may also want to further distinguish in-domain scores based on sub-domain (e.g., flights, hotels, events). Perhaps most importantly, we still need to conduct a full-scale evaluation with the above metric with several scorers and speaker pairs across different versions of the system to be able to provide actual results.

7 Acknowledgements

I would like to thank my advisor Lori Levin, Alon Lavie, Monika Woszczyna, and Aleksandra Slavkovic for their help and suggestions with this work.

References

M.Danieli and E.Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceeedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39.

L.Hirschman, D.Dahl, D.P.McKay, L.M.Norton, M.C.Linebarger. 1990. Beyond class A: A proposal for automatic evaluation of discourse. In *Proceedings of the Speech and Natural Language Workshop*, pages 109–113.

L.Hirschman and C.Pao. 1993. The cost of errors in a spoken language system. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pages 1419–1422.

J.Polifroni, L.Hirschman, S.Seneff, and V.Zue. 1992. Experiments in evaluating interactive spoken language systems. In *Proceedings of the DARPA Speech and NL Workshop*, pages 28–31.

E.Shriberg, E.Wade, and P.Price. 1992. Human-machine problem solving using spoken language systems (sls): Factors affecting performance and user satisfaction. In *Proceedings of the DARPA Speech and NL Workshop*, pages 49–54.

T.Sikorski and J.Allen. 1995. A task-based evaluation of the TRAINS-95 dialogue system. Technical report, University of Rochester.

A.Simpson, and N.A.Fraser. 1993. Black box and glass box evaluation of the SUNDIAL system. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pages 1423–1426.

M.Walker, D.J.Litman, C.A.Kamm, and A.Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. Technical Report TR 97.26.1, AT and T Technical Reports.

D.Gates, A.Lavie, L.Levin, A.Waibel, M.Gavalda, L.Mayfield, M.Woszczyna, P.Zhan. 1996. End-to-end Evaluation in JANUS: a Speech-to-speech Translation System. In *Proceedings of the 12th European Conference on Artificial Intelligence, Workshop on Dialogue, Budapest, Hungary.*