Evaluation of NLG in an end-to-end Spoken dialogue system- is it worth it?

Helen Hastie[1], Heriberto Cuayáhuitl[1], Nina Dethlefs[2], Simon Keizer[1], Xingkun Liu[1]

Abstract In the past 10 years, only around 15% of published conference papers include some kind of extrinsic evaluation of an NLG component in an end-to-end system. These types of evaluations are costly to set-up and run, so is it worth it? Is there anything to be gained over and above intrinsic quality measures obtained in off-line experiments? In this paper, we describe a case study of evaluating two variants of an NLG surface realiser and show that there are significant differences in both extrinsic measures and intrinsic measures. These significant differences would need to be factored into future iterations of the component and therefore, we conclude that extrinsic evaluations are worthwhile.

1 Introduction

Extrinsic evaluations of Spoken Dialogue System output components (NLG and TTS) are relatively rare. [1] surveyed published conference papers for NLG systems and found only 15% included some type of extrinsic evaluation in an end-to-end system. Similar published TTS studies are even rarer. Extrinsic evaluations that include testing end-to-end systems are highly labour-intensive to set up and cost on average more per data-point to collect due to the time taken to complete a whole dialogue rather than, for example, an off-line rating of written output. Input components (ASR/SLU) rely less on subjective measures of quality than output components and one is easily able to obtain intrinsic quality measures, such as WER, as well as extrinsic measures during end-to-end system evaluations. The question arises, therefore, whether it is worth running extrinsic evaluations for output components and whether the user can perceive any differences in quality whilst performing a task through dialogue. One could hypothesise that with all that is going on in a some-

[1]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, Scotland, e-mail: h.hastie, h.cuayahuitl, s.keizer, x.liu@hw.ac.uk · [2]Department of Modern Languages, University of Hull, England, e-mail: n.dethlefs@hull.ac.uk

times complex dialogue, that nuances in the style and naturalness of the output will not be perceptible by the user.

Here, we adopt the categorisation and terminology defined in [2] where the authors define *extrinsic measures* as those that assess the effect of a system on something external to it. This metric is further broken down into *user-task-success* (e.g. finding a restaurant) and *system purpose success* (e.g. changing the user's attitude towards the environment); here, we concentrate on the former. *Intrinsic measures* assess properties of systems or components in their own right, e.g. compared to some gold standard. [2] break these intrinsic measures down further into *output quality measures* (either automatic or hand-labelled) and *user like measures* (the subjective assessment of quality).

Previous end-to-end evaluations reported in the literature have had mixed results with regards evaluating NLG components embedded in a spoken dialogue system. One study does show differences in both *intrinsic* and *extrinsic* measures, with [3] showing significant differences in *extrinsic user-task-success* measures and one *intrinsic user like measure* on voice quality when testing adaptive vs non-adaptive NLG. For the ILEX project, [4] evaluated two versions of NLG for museum artifact text generation and found significant differences in *intrinsic user like measures* but not for *extrinsic measures*. For the M-PIRO system [5], a follow-on from ILEX in the same domain, the authors were interestingly able to show significant improvement on *extrinsic user-task-success measures* of comprehension accuracy and *extrinsic user-assessed learning gain* but not *intrinsic user like measures* such as 'interestingness' and enjoyment.

This work follows on from [6], where we showed through *intrinsic* measures that no differences are perceivable between an NLG trained on automatically labelled data (Proposed) and an NLG trained on hand-labelled data (TopBound). These measures include: understanding, phrasing and naturalness. Note, there was, however, a significant difference between these two systems and hand-crafted output. These two variants of the NLG component were integrated into an end-to-end system and an evaluation of comparable size, in terms of generated utterances, is reported here.

2 Evaluating the NLG component in an end-to-end system

Our domain is an interactive system that provides restaurant recommendations to users with varying preferences and constraints. We are evaluating two statistical realisers: one trained on labelled data (TopBound) and one on unlabelled data (Proposed) in the hope of increasing the portability of the surface realiser to new domains. In the first stage, automatic semantic labelling is applied to unlabelled utterances (see [6]). In the second stage, the automatically labelled data is then used to train an existing statistical surface realiser [7], which uses Conditional Random Fields (CRFs) to generate outputs from input dialogue acts. Our semantic labelling method is based on unsupervised clustering of clauses found in unlabelled input data according to their lexical and semantic similarity. The underlying hypothesis is

that the more similar clauses are, in terms of their lexical and semantic properties, the more likely they are to represent the same semantic slot, e.g. *kidsAllowed*, or *goodForMeal*. A component evaluation reported earlier [6] confirmed that this automatic labelling technique can attain good clustering accuracy results. Examples from the two variants of the NLG surface realiser are: **TopBound:** "Right in the heart of central Richmond lies Kirin Chinese restaurant, a well-established neighbourhood favourite."; and **Proposed:** "Right in the heart of central Richmond lies the well-established neighbourhood favourite of Kirin Chinese restaurant."

The statistical surface realiser was integrated as part of the PARLANCE dialogue system [8][1]. The system architecture includes the following components: the ATK speech recogniser [9]; an SLU dependency parser with unsupervised word embeddings [10]; an Interaction Manager that uses Gaussian Process reinforcement learning with a policy trained on the top-bound ontology [11]; and finally, the generated outputs are given as input to the TTS engine described in [12].

A task-based evaluation was conducted with workers recruited via Crowdflower[2]. The workers were asked to call the system and find restaurants in certain areas of San Francisco (U.S.A.) according to certain predefined scenarios, e.g. "You want to find a restaurant in the center and it should serve Indian food. If there is no such venue how about one with African food? You want to know the address and whether it is good for lunch." 664 dialogues were collected from 72 participants. Participants were paid $2.00 on completion of four dialogues. After the participants have completed a dialogue, they were given 5 subjective questions where Q1 was a binary Yes/No for perceived *user-task-success* and Q2-5 were on a 6-point rating scale and cover a variety of aspects of dialogue (see Table 2).

Evaluation mode:		**Intrinsic output quality**			**Extrinsic**	
System	**Num of Dialogues**	**Num of Turns**	**Length (sec)**	**Avg wds per turn**	**TS**	**SubjTS**
TopBound	365	15.23	10.51	12.32	60.82%	94.79%*
Proposed	299	14.71	10.23	11.89*	61.20%	89.63%

Table 1 *Extrinsic user-task-success (TS), subjective user-task-success (SubjTS)* and various *intrinsic output quality measures* capturing dialogue length. * indicates $p < 0.05$ using a χ^2 test for TS/SubjTS and 1-way unpaired t-test for length metrics

Evaluation mode:	**Intrinsic user like**			
System	**InfoFound**	**Understanding**	**InfoPresentation**	**Repetitiveness**
TopBound	4.70(6)	4.54(6)	4.56(6)*	4.36(6)
Proposed	4.60(6)	4.31(6)	4.31(6)	4.22(6)

Table 2 *Intrinsic user like measures* from the post-questionnaire on a 6-point rating scale for the mean (mode).* indicates p$<$ 0.05 for a Mann-Whitney U test

[1] Note, the data resources referred to in this paper are available at http://www.parlance-project.eu

[2] http://crowdflower.com

As seen in Table 1, the TopBound system is *perceived* as significantly more successful than the Proposed system in retrieving a relevant restaurant (SubjTS). However, in terms of actual hand-annotated task success[3] (TS), there is no significant difference. The difference in TS and SubjTS, we believe, is due to participants overestimating their success, i.e. if they got information on *any* restaurant they marked it as a success. The Proposed system has significantly shorter turns in terms of average words per turn (an *intrinsic automatic output quality measure*), which may be due to missing or confused slot information.

As seen in Table 2, the only significant difference between the *intrinsic user like measures* was for Information Presentation. Therefore, even though the outputs are relatively similar, the user is still able to perceive a difference in the Information Presentation category, thus highlighting areas for improvement. In [7], we show through *user like measures* that, by their very nature, using CRFs for surface realisation results in utterances that are *less repetitive* than baseline systems. This still holds when trained on automatically labelled data, as there is little perceived difference in terms of repetitiveness between the TopBound and Proposed systems.

Error analysis reveals that in approximately 4% of the cases, the CRF trained on automatically labelled data realises somewhat anomalous utterances such as "The Kirin restaurant is a perfect place for children out". This may contribute to the decrease in subjective evaluation scores. A further aspect revealed in our error analysis is that generated outputs can occasionally contain segments of information that are not part of the original semantic input form. An example is the realisation "Right in the heart of central Richmond lies the well-established neighbourhood favourite of Kirin Chinese restaurant." for semantic input form *inform(restaurant, area="central Richmond", venueName="Kirin", foodType=Chinese)*. The fact that the restaurant is "a well-established neighbourhood favourite" is not derived from the knowledge base but rather constitutes an artifact of the training data. End-to-end evaluations, particularly those "in the wild" with users actually visiting the recommended restaurants, may show this to be a false statement; again this would not be evident in isolated utterances evaluated off-line.

Future work will transfer what we have learned in this evaluation to new domains, going beyond restaurant recommendations [13]. In addition, this work supports the argument for joint optimisation of NLG with other components such as the Interaction Manager. The study described here is not a typical evaluation where one has a Proposed system that one is claiming to be better than an existing technique or baseline system. Rather, we aim to build an automatic system (Proposed) in the hope that it would perform *as well as* a version that involves costly development in terms of hand-labelling data (TopBound). Where previous studies using end-to-end systems have shown mixed results, mostly getting significant differences in either *intrinsic* or *extrinsic* but rarely both, we have indeed shown that even nuanced differences in NLG are perceptible during interaction, while at the same time influencing perceived task success.

[3] *Extrinsic user-task-success* was hand-annotated by a single annotator, being set to 1 if the caller received information on a restaurant that matched their request and if other information (e.g. *address, name, phoneNumber*) was asked for and correctly received.

Acknowledgements

This research was funded by the European Commission FP7 programme FP7/2011-14 under grant agreement no. 287615 (Parlance). We thank all members of the PARLANCE consortium for their help in designing, building and testing the Parlance end-to-end spoken dialogue system. We would also like to acknowledge other members of the Heriot-Watt Parlance team in particular Prof. Oliver Lemon and Verena Rieser.

References

1. Gkatzia, D., Mahamood, S.: A snapshot of NLG evaluation practices 2005 to 2014. In: Proceedings of ENLG. (2015)
2. Belz, A., Hastie, H.: Towards comparative evaluation and shared tasks for NLG in interactive systems. In Bangalore, S., Stent, A., eds.: Natural Language Generation in Interactive Systems. Cambridge University Press, Cambridge (2014) 302–350
3. Rieser, V., Lemon, O., Keizer, S.: Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. IEEE/ACM Transactions on Transactions on Audio, Speech, and Language Processing **22**(5) (May 2014)
4. Cox, R., O'Donnell, M., Oberlander, J.: Dynamic versus static hypermedia in museum education: An evaluation of ILEX, the intelligent labelling explorer. In: Proceedings of AIED. (1999)
5. Karasimos, A., Isard, A.: Multi-lingual evaluation of a natural language generation systems. In: Proceedings of LREC. (2004)
6. Cuayáhuitl, H., Dethlefs, N., Hastie, H., Liu, X.: Training a Statistical Surface Realiser from Automatic Slot Labelling. In: Proceedings of SLT, South Lake Tahoe, CA, USA (2014)
7. Dethlefs, N., Hastie, H., Cuayáhuitl, H., Lemon, O.: Conditional random fields for responsive surface realisation using global features. In: Proceedings of ACL. (2013)
8. Hastie, H., Aufaure, M.A., Alexopoulos, P., Cuayáhuitl, H., Dethlefs, N., Gašić, M., Henderson, J., Lemon, O., Liu, X., Mika, P., Ben Mustapha, N., Rieser, V., Thomson, B., Tsiakoulis, P., Vanrompay, Y., Villazon-Terrazas, B.: Demonstration of the PARLANCE system: a data-driven incremental, spoken dialogue system for interactive search. In: Proceedings of SIGDIAL. (2013)
9. Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P.: The HTK Book Version 3.0, Cambridge University, UK (2000)
10. Yazdani, M., Breslin, C., Tsiakoulis, P., Young, S., Henderson, J.: Domain adaptation in ASR and SLU. Technical report, PARLANCE FP7 Project (2014)
11. Gašić, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., Young, S.: POMDP-based dialogue manager adaptation to extended domains. In: Proceedings of SIGDIAL. (2013)
12. Tsiakoulis, P., Breslin, C., Gašić, M., Henderson, M., Kim, D., Young, S.J.: Dialogue context sensitive speech synthesis using factorized decision trees. In: Proceedings of INTERSPEECH. (2014)
13. Cuayáhuitl, H., Dethlefs, N., Hastie, H.: A Semi-Supervised Clustering Approach for Semantic Slot Labelling. In: Proceedings of ICMLA, Detroit, MI, USA (2014)