Getting to Know Users: Accounting for the Variability in User Ratings

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser and Oliver Lemon
Heriot-Watt University
Mathematical and Computer Sciences
Edinburgh EH14 4AS, UK
n.s.dethlefs@hw.ac.uk

Abstract

Evaluations of dialogue systems and language generators often rely on subjective user ratings to assess output quality and performance. Humans however vary in their preferences so that estimating an accurate prediction model is difficult. Using a method that clusters utterances based on their linguistic features and ratings (Dethlefs et al., 2014), we discuss the possibility of obtaining user feedback implicitly during an interaction. This approach promises better predictions of user preferences through continuous re-estimation.

1 Introduction

Given the subjective nature of human language, many evaluation studies in dialogue systems and natural language generation rely on subjective user ratings to assess performance and acceptability. A shared problem however is that humans vary considerably in their individual preferences, making it difficult to estimate an accurate prediction model. To account for individual preferences and still make accurate predictions, in Dethlefs et al. (2014) we proposed to cluster utterances based on their linguistic properties and the ratings they receive from groups of individual users. Results confirmed that prediction accuracy improves significantly in this way: predictive models based on clusters of ratings lead to significantly better predictions than models based on an average population of ratings–as is currently state of the art.

The required clusters can be obtained from minimal information about an individuals user's preferences, such as a single user rating alone. One drawback of our method so far, however, is that it remains unclear how user ratings can best be obtained during an ongoing human-computer interaction. Requesting ratings explicitly may be the easiest way, but can disrupt interactions. Here, we discuss alternatives based on (a) the interaction history, (b) interactive alignment, and (c) multimodal information. We discuss the potential of each of these ideas to implicitly elicit user feedback on system utterances during an interaction.

2 State of the Art

The problem of variability in subjective user ratings has been recognised by various authors in different domains such as recommender systems (O'Mahony et al., 2006; Amatriain et al., 2009), sentiment analysis (Pang and Lee, 2005), content selection (Jordan and Walker, 2005; Dale and Viethen, 2009) and surface realisation (Walker et al., 2007; Dethlefs et al., 2014). The primary method of capturing individual differences in statistical models so far has been to train separate models for individual users (Dale and Viethen, 2009; Walker et al., 2007). In practice, this can often be done by including the user's ID as a feature for classification or regression. This tends to significantly improve performance for the user in question, but fails to generalise to users with no prior ratings. We can therefore distinguish (a) systems that estimate prediction models from an average population of users–and thereby ignore the existing variability; and (b) systems that are trained for individual users and fail to generalise to unseen instances.

3 Using Clustering to Account for Variable User Ratings

In Dethlefs et al. (2014), we have presented an approach that aims to find a middle ground between making predictions from an average population of users and training an individual model for each new user. Figure 1 provides an illustration of the approach. In essence, the idea is to learn a mapping between the linguistic features of a group of utterances that receive similar ratings, e.g. ratings

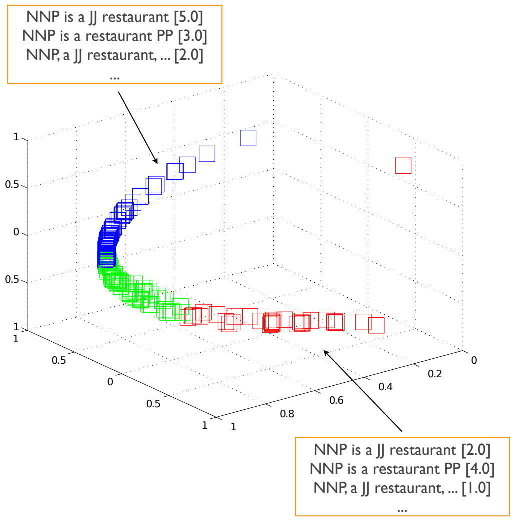

Figure 1: Clusters are estimated based on linguistic features and ratings. Prediction is then aided by estimating which cluster a new user might rate according to. Users in the same cluster (indicated in different colours) tend to rate utterances similarly.

for *politeness* on a scale of 1-5. We used multiple multivariate regression and features included lexical information, such as the presence of individual words, the average tf-idf score of an utterance, and syntactic features such as the depth of syntactic embedding. Clusters are identified from pair-wise similarities between data points using the Kullback-Leibler divergence (Cuayáhuitl et al., 2005). A spectral clustering algorithm performs dimensionality reduction and clusters similar pairs of linguistic features and user ratings into the same cluster and dissimilar pairs into separate clusters. Results have shown that minimal information on user preferences is sufficient to perform significantly better than based on an average population of users. Please see Dethlefs et al. (2014) for details on the approach and an evaluation.

4 Discussion

This section discusses three possible options of obtaining user feedback during an interaction.

Interaction Context including dialogue moves that follow a system utterance or incremental phenomena such as barge-ins or backchannels can all offer insights into a user's perception of an ongoing interaction (Janarthanam and Lemon, 2014). For example, barge-ins and unforeseen dialogue moves can be indicative of a problematic dialogue, whereas backchannelling and alignment with the system can indicate success. Based on this, a possibility is to extend the PARADISE framework (Walker et al., 1997) by estimating a regression model that predicts user ratings based on incremental dialogue phenomena in an online fashion. However, it is likely that such phenomena also exhibit variation between individual users. They can therefore provide feedback on subjective as well as objective evaluation scales.

Interactive Alignment could be applied under the hypothesis that adapting to the linguistic features found in users' speech would have a favourable influence on their perception of the system and lead to positive ratings. This assumption is based on psycholinguistic evidence that humans prefer to interact with humans that align with them (Levelt and Kelter, 1982). Further, computational studies have shown that interactive alignment in human-computer interaction can be created and recognised by users (Brockmann et al., 2005; Isard et al., 2006; Dethlefs, 2013). In our case, results of the ASR could be analysed and linguistic features extracted. An experimental study would have to confirm that such alignment is plausible, noticeable to users and perceived positively.

Multimodal Information could provide valuable feedback cues, including user hesitations and pauses or even gesture recognition or eye-tracking. Ultimately, our goal is to use non-verbal cues as feedback signals in an interaction so that system behaviour can be continuously re-estimated and improved (Cuayáhuitl and Dethlefs, 2011). Perceptive cues such as the user frowning, losing attention, or hesitating regarding the next step to take in the interaction could indicate problems in the interaction, while smiling or continued attention could be interpreted as positive cues. A data collection and analysis would need to explore the full range of multimodal cues available.

Future work will explore these ideas and analyse their practical advantages and drawbacks. To do this, we will use the PARLANCE system, a data-driven, incremental and spoken interactive system (Hastie et al., 2013), which also exists as a mobile app (Hastie et al., 2014). Implicit feedback elicitations could thus be combined with explicit feedback to gain more information on users and allow the personalisation of system output.

Acknowledgments

This research was funded by the EC FP7 programme FP7/2011-14 under grant agreement no. 287615 (PARLANCE).

References

Xavier Amatriain, Josep M. Pujol, and Nuria Oliver. 2009. I like It... I Like It Not: Evaluating User Ratings Noise in Recommender Systems. In *In the 17th International Conference on User Modelling, Adaptation, and Personalisation (UMAP)*, pages 247–258, Trento, Italy. Springer-Verlag.

Carsten Brockmann, Amy Isard, Jon Oberlander, and Michael White. 2005. Modelling alignment for affective dialogue. In *Proceedings of the UM-05 Workshop on Adapting the Interaction Style to Affective Factors*.

Heriberto Cuayáhuitl and Nina Dethlefs. 2011. Optimizing Situated Dialogue Management in Unknown Environments. In *Proceedings of INTERSPEECH*, pages 1009–1012.

Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2005. Human-Computer Dialogue Simulation Using Hidden Markov Models. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU)*, San Juan, Puerto Rico.

Robert Dale and Jette Viethen. 2009. Referring Expression Generation Through Attribute-Based Heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG)*, Athens, Greece.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based Prediction of User Ratings for Stylistic Surface Realisation. In *Proceedings of the European Chapter of the Annual Meeting of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden.

Nina Dethlefs. 2013. *Hierarchical Joint Learning for Natural Language Generation*. PhD Thesis, University of Bremen, Germany.

Helen Hastie, Marie-Aude Aufaure, Panos Alexopoulos, Heriberto Cuayhuitl, Nina Dethlefs, James Henderson Milica Gasic, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha, Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay, Boris Villazon-Terrazas, and Steve Young. 2013. Demonstration of the PARLANCE System: A Data-Driven, Incremental, Spoken Dialogue System for Interactive Search. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*.

Helen Hastie, Marie-Aude Aufaure, Panos Alexopoulos, Hughes Bouchard, Heriberto Cuayáhuitl, Nina Dethlefs, Milica Gasic, James Henderson, Oliver Lemon, Xingkun Liu, Peter Mika, Nesrine Ben Mustapha, Tim Potter, Verena Rieser, Blaise Thomson, Pirros Tsiakoulis, Yves Vanrompay, Boris Villa-Terrazas, Majid Yazdani, Steve Young, and Yanchao Yu. 2014. The PARLANCE Mobile App for Interactive Search in English and Mandarin. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*.

Amy Isard, Carsten Brockmann, and Jon Oberlander. 2006. Individuality and Alignment in Generated Dialogues. In *Proceedings of the 4th International Natural Language Generation Conference (INLG)*, Sydney, Australia.

Srini Janarthanam and Oliver Lemon. 2014. Adaptive generation in dialogue systems using dynamic user modeling. *Computational Linguistics*. (in press).

Pamela Jordan and Marilyn Walker. 2005. Learning Content Selection Rules for Generating Object Descriptions in Dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

Willem Levelt and Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive Psychology*, 14:78–106.

Michael O'Mahony, Neil Hurley, and Guénolé Silvestre. 2006. Detecting Noise in Recommender System Databases. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)s*. ACM Press.

Bo Pang and Lillian Lee. 2005. Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marilyn Walker, Diane Litman, Candace Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL)*, Madrid, Spain.

Marilyn Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and Domain Adaptation in Sentence Planning for Dialogue. *Journal of Artificial Intelligence Research*, 30(1):413–456.