Using Collaborative Filtering to Predict User Utterances in Dialogue

Ryuichiro Higashinaka, Noriaki Kawamae, Kohji Dohsaka, and Hideki Isozaki

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, 619-0237 Kyoto, Japan.
{rh,kawamae.noriaki,dohsaka,isozaki}@cslab.kecl.ntt.co.jp

Abstract. This paper proposes using collaborative filtering, a technique for using other users' information to model the behavior of a certain user, to predict users' evaluative expressions for entities in dialogue. Previous studies have found that inducing users' empathic utterances towards systems can improve user satisfaction. Predicting what users may utter and communicating this information in advance would make it easy for users to show empathy, leading to possible improvement in the quality of dialogue. Experimental results show that our approach, which uses the similarity users and entities, can significantly improve the prediction of evaluative expressions compared to a baseline that ignores such similarity.

1 Introduction

Although task-oriented dialogue systems have been actively studied over the years [M. Walker et al., 2002], systems that aim to affect the minds of users are beginning to be actively investigated [Bickmore and Picard, 2005,Dohsaka et al., 2009]. In our previous work, which investigated the effects of self-disclosure and empathy on closeness and user satisfaction, we found that inducing emphatic utterances of users is most important for improving closeness and user satisfaction [Higashinaka et al., 2008].

This paper examines how such empathic utterances can be induced using collaborative filtering, a technique for using other users' information to model the behavior of a certain user [Breese et al., 1998]. We adopt this technique to predict user utterances (in particular, evaluative expressions) in dialogue. For example, for users who think cats are capricious, if the system could predict this and utter, "Cats are capricious, aren't they?" in advance, users would likely to agree and may feel close to the system.

Collaborative filtering has already been used to predict the ratings of user reviews [Amatriain et al., 2009,Titov and McDonald, 2008]. However, it has never been used for predicting expressions of users in dialogue. Recently, automatically mining emphatic utterances from the web has also been proposed [Shimizu et al., 2009]. However, this does not take into account the information of users.

2 Approach

For a user U_i $(1 \leq i \leq n)$ and an entity E_j $(1 \leq j \leq m)$, we want to predict the expressions $\{e_1 \ldots e_k\}$ for that entity from other users' information. Since the work is still preliminary, we focus on predicting a user's evaluative expressions (mainly adjectives) in this paper, although we plan to extend our approach to dealing with more complex expressions.

First, we make a frequency table where rows represent users and columns represent evaluative expressions, showing the use of evaluative expressions by users. Here, we weight the frequencies by Residual IDF (RIDF). From this table, using a variant of the cosine similarity metric [Amatriain et al., 2009], we calculate the user similarity between U_i and U_j by

$$\mathrm{sim}(U_i, U_j) = \frac{\boldsymbol{u_i} \cdot \boldsymbol{u_j}}{\mid \boldsymbol{u_i} \mid\mid \boldsymbol{u_j} \mid} \cdot \frac{2N_{i \cup j}}{N_i + N_j} \tag{1}$$

where $\boldsymbol{u_i}$ and $\boldsymbol{u_j}$ mean the weighted frequency vectors for U_i and U_j, N_i the number of entities mentioned by U_i, and $N_{i \cup j}$ the number of entities mentioned by both U_i and U_j.

Then, we make another frequency table where rows represent entities and columns represent evaluative expressions, showing how entities are described by users using the evaluative expressions. Here, each cell represents the number of times an evaluative expression e_k was used for an entity by all users except for U_i. We calculate this count $\mathrm{entity}(U_i, E_l, e_k)$ by

$$\mathrm{entity}(U_i, E_l, e_k) = \sum_{h=1}^{n} \mathrm{sim}(U_i, U_h) \cdot \mathrm{freq}(U_h, E_l, e_k) \tag{2}$$

where freq is a function that counts how many times a user U_h used an evaluative expression e_k for an entity E_l. The frequency is weighted by the similarity between U_i and U_j so that the counts of users who are not similar to U_i can be treated lightly. This way, each row can represent a frequency vector of evaluative expressions adapted to U_i. In addition, since similar entities may share similar evaluative expressions, we update the counts by

$$\mathrm{entity}_{\mathrm{update}}(U_i, E_j, e_k) = \sum_{l=1}^{m} \mathrm{sim}(E_j, E_l) \cdot \mathrm{entity}(U_i, E_l, e_k) \tag{3}$$

Finally, when predicting the evaluative expressions of U_i for E_j, we rank the expressions $\{e_1 \ldots e_k\}$ in descending order of $\mathrm{entity}_{\mathrm{update}}(U_i, E_j, e_k)$.

3 Experiment

We prepared sets of evaluative expressions for 90 animals (i.e., entities) for 50 users. We created this data set from our text-chat dialogue data in the animal discussion domain, in which a user and a system discuss their favorite animals (see details in [Higashinaka et al., 2008]).

We extracted the users' evaluative expressions for the animals from the dialogue act annotation we performed on user utterances. Here, the evaluative expressions are adjectives, adjective verbs, and phrases that match the pattern

Table 1. Top-3 accuracies for evaluative expressions averaged over all users depending on the skewness threshold t. Asterisks indicate statistical significance (p<0.05) over Baseline. The numbers in parentheses denote the number of animals over t.

	none (90)	t=1.0 (74)	t=1.5 (59)	t=2.0 (35)	t=2.5 (17)
Baseline	0.753	0.760	0.756	0.771	0.776
UserSim	0.752	0.765	0.774*	0.794*	0.799
AnimalSim	0.733	0.741	0.759	0.773	0.802*
UserSim+AnimalSim	0.740	0.740	0.755	0.760	0.778

[`NN is ADJ`] where `NN` and `ADJ` stand for a general noun and an adjective, respectively. For example, we have "*clever*" and "*eyes are cute*" for dolphins. Note that, in extracting the expressions, we did not select expressions that occurred less than ten times in the data because such expressions may be too specific to certain users. We extracted 48 evaluative expressions in all.

For the prediction experiment, we excluded one user's evaluative expressions for an animal from the data set and predicted them using our approach. This process was repeated for all animals for all users in a round robin fashion. For evaluation, we calculated the accuracy of evaluative expressions by first predicting top-N evaluative expressions for the animals, and then calculating the ratio of animals for which the top-N expressions contained those actually uttered by the user. We used 3 for N in this experiment. We prepared three configurations of our proposed approach and compared them with a baseline. They are as follows.

UserSim The evaluative expressions are simply ranked by Eqn.2.
AnimalSim The evaluative expressions are ranked by Eqn.3 but the similarity of users is not used; namely, Eqn.1 returns 1.
UserSim+AnimalSim The evaluative expressions are ranked by Eqn.3.
Baseline The evaluative expressions are ranked by Eqn.2 and Eqn.1 returns 1.

Table 1 shows the top-3 accuracies for the evaluative expressions averaged over all users. We selected the animals for evaluation on the basis of the *skewness* of the distribution of evaluative expressions because we noticed that the ease of prediction depends on whether an animal is expressed by only a few dominating evaluative expressions. The skewness is calculated by $\{N/(N-1)/(N-2)\}\sum_{i=1}^{N}(x_i-\mu)^3/\sigma^3$, where N, x_i, μ, and σ denote the number of evaluative expressions, the frequency of each evaluative expression, the mean, and the standard deviation of the frequencies of the evaluative expressions for a given animal, respectively. Since the skewness can capture such distortion in a distribution, we used the skewness threshold t to remove certain animals from evaluation. For the statistical comparison of the accuracies, we performed a sign test that compares the number of users whose accuracies were higher or lower than the baseline.

From the table, it can be seen that UserSim significantly outperformed the baseline when $t = 1.5$ (p=0.019) and $t = 2.0$ (p=0.013). AnimalSim also outperformed the baseline when $t = 2.5$ (p=0.035). Figure 1 shows how the accuracies change depending on t. Although our current approach does not work sufficiently for animals with low skewness (i.e., animals that are expressed by many low-frequency evaluative expressions), the results demonstrate the effectiveness of using other users and entities for the prediction of users' evaluative expressions.

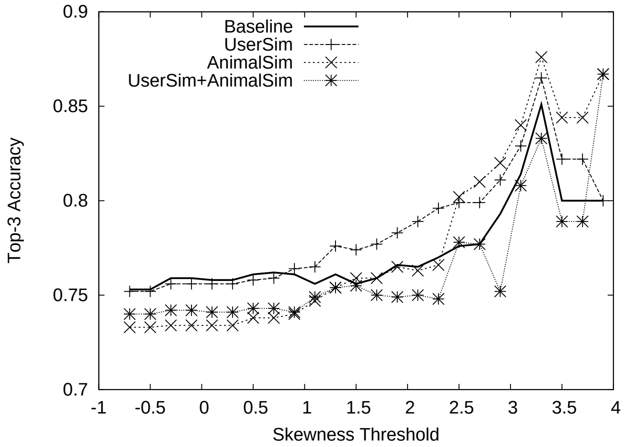

Fig. 1. Plot of top-3 accuracies for evaluative expressions averaged over all users depending on the skewness threshold t.

4 Summary and Future Work

This paper proposed using collaborative filtering to predict users' evaluative expressions for entities in dialogue. The experimental results show that our approach is promising. As future work, we plan to improve the prediction accuracy and coverage of entities. We also plan to use larger and more realistic data to verify our approach. We also need to show that the effectiveness of our approach in an ongoing dialogue. Finally, we also plan to apply our approach to spoken dialogue data.

References

[Amatriain et al., 2009] Amatriain, X., Lathia, N., Pujol, J., Kwak, H., and Oliver, N. (2009). The Wisdom of the Few: A Collaborative Filtering Approach Based on Expert Opinions from the Web. In *Proc. SIGIR*, pages 532–539.

[Bickmore and Picard, 2005] Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

[Breese et al., 1998] Breese, J. S., Heckerman, D., and Kadie, C. M. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. the 14th Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52.

[Dohsaka et al., 2009] Dohsaka, K., Asai, R., Higashinaka, R., Minami, Y., and Maeda, E. (2009). Effects of conversational agents on human communication in thought-evoking multi-party dialogues. In *Proc. SIGDIAL*, pages 217–224.

[Higashinaka et al., 2008] Higashinaka, R., Dohsaka, K., and Isozaki, H. (2008). Effects of self-disclosure and empathy in human-computer dialogue. In *Proc. SLT*, pages 109–112.

[M. Walker et al., 2002] M. Walker et al. (2002). DARPA Communicator evaluation: Progress from 2000 to 2001. In *Proc. ICSLP*, pages 273–276.

[Shimizu et al., 2009] Shimizu, T., Inui, K., and Matsumoto, Y. (2009). Back-channel feedback using evaluation expression for chat. *SIG-SLUD-A803-02, Japanese Society of AI*, pages 7–12. (in Japanese).

[Titov and McDonald, 2008] Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *Proc. WWW*, pages 111–120.