Automatic Evaluation of Chat-oriented Dialogue Systems using Large-scale Multi-references

Hiroaki Sugiyama, Toyomi Meguro and Ryuichiro Higashinaka

Abstract The automatic evaluation of chat-oriented dialogue systems remains an open problem. Most studies have evaluated them by hand, but this approach requires huge cost. We propose a regression-based automatic evaluation method that evaluates the utterances generated by chat-oriented dialogue systems based on the similarities to many reference sentences and their annotated evaluation values. Our proposed method estimates the scores of utterances with high correlations to the human annotated scores; the sentence-wise correlation coefficients reached 0.514, and the system-wise correlation were 0.772.

1 Introduction

The enormous cost of evaluating chat-oriented dialogue systems is one major obstacle to improve them. Previous work has evaluated dialogue systems by hand [1, 2], which is a common practice in dialogue research. However, such an approach not only requires a huge cost but it is also not replicable; i.e., it is difficult to compare a proposed system's scores with the previously reported scores of other systems.

As a first trial of substituting human annotations, Ritter et al. introduced BLEU, which is a reference-based automatic evaluation method widely used in the assessment of machine-translation systems [3, 4]. They evaluate their dialogue systems on the basis of the appropriateness of each one-turn response for input sentences instead of whole dialogues. While such a reference-based evaluation methodology shows high correlations with human annotators in machine-translation, they reported that

Hiroaki Sugiyama
NTT Communication Science Labs., Kyoto, Japan, e-mail: sugiyama.hiroaki@lab.ntt.co.jp

Toyomi Meguro
NTT Communication Science Labs., Kyoto, Japan, e-mail: meguro.toyomi@lab.ntt.co.jp

Ryuichiro Higashinaka
NTT Media Intelligence Labs., Kanagawa, Japan, e-mail: higashinaka.ryuichiro@lab.ntt.co.jp

the reference-based approach fails to show high correlation with human annotations in the evaluation of chat-oriented dialogues. In machine-translation, since systems are required to generate sentences that have exactly the same meaning as the original input sentences, the appropriate range of the system outputs is so narrow that only one or just a few reference sentences are enough to cover them. On the other hand, in chat-oriented dialogues, since the appropriate range is likely to be much larger than in machine-translation, such a small number of references is likely to be insufficient.

Galley et al. proposed Discriminative BLEU (Δ BLEU), which leverages 15 references with manually annotated evaluation scores to estimate the evaluation of chat-oriented dialogue system responses [5]. Their method leverages *negative* references in addition to customary *positive* references in the calculation of BLEU and evaluates sentences that resemble negative references as inappropriate. This increases the correlation with human judgment up to 0.48 of Pearson's r when the correlation is calculated with 100 sentences as a unit; however, they also reported that sentence-wise correlation remained low: $r \leq 0.1$. The reason is probably that their method evaluates a sentence that is far from all the references as neutral rather than inappropriate.

We propose a *regression*-based approach that automatically evaluates chat-oriented dialogue systems by leveraging the distances between system utterances and a large number of positive and negative references. We expect our regression-based approach to appropriately evaluate sentences that are not similar to all of the references. We also gathered a larger scale of references than Galley's work and examined the effectiveness of the number of references over the estimation performance.

2 Multi-reference-based evaluation

This section explains how we gather positive/negative sentences by humans, consistently evaluate them among the annotators, and automatically estimate their evaluation scores.

2.1 Development of reference corpus

We developed a multi-reference corpus that contains both positive and negative reference sentences (responses to input sentences). To collect reasonable input-response pairs for automatic evaluation, first we collected utterance-like sentences as input sentences from the web and real dialogues between humans. To remove sentences that require understanding of the original contexts, human annotators rated the *comprehensibility scores* of the collected sentences (degrees of how the annotators can easily understand the situations of the sentences), and we randomly chose sentences with high comprehensibility scores.

After collecting the input sentences, 10 non-expert reference writers created response sentences that would satisfy users. To intentionally gather inappropriate responses, we designed the following two constraints of their creation: *character-length limitation* and *masked input sentences*. The character-length limitation, which narrows the available expressions, decreases the naturalness of the references. The limitation on masked input sentences means that the reference writers create responses using sentences whose words are partly deleted. For example, when we mask 60% of the words in the following sentence *What is your favorite subject?*, *What is *** *** ***?* or **** is *** favorite ***?* is shown to the reference writers. This enables us to gather response sentences that have irrelevant content to the original input sentences and simultaneously maintain the syntactic naturalness of the responses. In addition, to add other types of inappropriate sentences to the negative references, we gathered sentences that were generated by existing dialogue systems described in Section 3.1.2. .

2.2 Evaluation of references

Human annotators evaluate reference sentences in terms of their naturalness as responses. In this work, we adopted the **pairwise winning rate** over all other references as an evaluation score of a reference sentence. If a sentence is judged to be more natural than all the other references, its evaluation score is 1; a sentence that is judged the least natural obtains an evaluation score of 0. Our preliminary experiment showed that if the evaluation scores are rated on a 7-point Likert scale, they tend to be either maximum or minimum; 45% were rated as 7 and 25% as 1. Hence it is difficult to determine the differences among the references by their scores. On the contrary, the winning rates vary broadly, and we can precisely distinguish differences among the references.

The drawback of the winning rate is its evaluation cost; the number of pairwise evaluations of N references is $N(N-1)/2$. However, pairwise evaluations for partly sampled pairs are reported to be satisfactorily accurate to maintain the winning rates [6].

2.3 Score estimation methods

Our method estimates a score of a pair of an input sentence and its response T. We considered the following three approaches in order to automatically evaluate system responses using the gathered pairs of input-references with human-annotated evaluation scores (pairwise winning rates).

Average of metrics (AM)

This method outputs an estimation score of target sentence T with the average of sentence-wise similarities $s_{(T,R_m)}$ with top-M similar reference sentences R_m as follows:

$$E_{AM}(T) = \frac{\Sigma_{m=1}^{M} s_{(T,R_m)}}{M}. \quad (1)$$

This utilizes only the similarities with the references and resembles the approach for machine-translation. Since this assumes that only positive references are input, we just use manually created references without a masking constraint.

Weighted scores (WS)

This method first calculates the sentence-wise similarities $s_{(T,R_m)}$ with top-M similar reference sentences R_m. This method also calculates the evaluated scores e_m (winning rates) of the top-M similar references. Then it outputs the average of the scores e_m of the top-N similar references weighted by the similarities $s_{(T,R_m)}$ as

$$E_{WS}(T) = \frac{\Sigma_{m=1}^{M} \{e_m \cdot s_{(T,R_m)}\}}{M}. \quad (2)$$

Regression

This estimates the evaluation scores with regression models like Support Vector Regression (SVR) [7]. We used similarity metrics $s(T,R_n)$ for N references as features and trained the model with data $\mathscr{D} = (\mathbf{x}_i, e_i)_{i=1}^{N}$, where $\mathbf{x}_i = \{s(R_i,R_j)\}, j \in \{1,\ldots,N\}$. Here, $s(R_i,R_j)$ means a similarity score between reference R_i and R_j. We developed a regression model for each input sentence.

3 Experiments

First we gathered pairs of input and reference sentences with evaluation scores. Then, based on the references, we developed evaluation score estimators and examined the effectiveness of our multi-reference approach.

3.1 Settings

3.1.1 Input sentences

We sampled input sentence candidates from a chat-oriented dialogue corpus as well as a Twitter corpus. Both corpora contain only Japanese sentences. The chat-oriented dialogue corpus consisted of 3680 one-to-one text-chat dialogues between Japanese speakers without specified topics [1]. From this corpus, we extracted input sentence candidates whose dialogue-acts were related to self-disclosure. From Twitter, we sampled sentence candidates that contain topic words, which were extracted from the top-10 ranked terms of Google trends in 2012 in Japan[1].

To remove candidates that require the contexts of the original dialogues to be understood by the writers, we recruited two annotators who rated the *comprehensibility scores* of the sentence candidates on a 5-point Likert scale and only used sentences that received scores of 5 from both annotators as input sentences. For the following experiment, we used ten input sentences: five randomly sampled from the conversational corpus and five from the Twitter corpus. The number of input sentences may be small due to the cost of labeling as we describe in the next section.

3.1.2 Reference sentences and evaluations

Method	$N_c < 50$	$10 \leq N_c < 50$	$N_c < 10$	Sum
Human (no mask)	18	18	6	42
Human (30% mask)	6	6	2	14
Human (60% mask)	6	6	2	14
IR-status	10	0	0	10
IR-response	10	0	0	10
Rule	10	0	0	10
Sum	60	30	10	100

Table 1 Statistics of gathered references for an input sentence. Human denotes number of manually created references and the others are automatically created references.

Using the ten selected input sentences, ten reference sentence writers (not the annotators for comprehensibility scores) created references. Each writer created seven reference sentences for each input sentence under two constraints: character-length limitation and masked input sentences. Table 1 shows the statistics of the gathered references. As a limitation of character length N_c, one reference writer created three sentences under the $N_c < 50$ condition (nearly free condition) that only limits excessively long references, three sentences under $10 \leq N_c < 50$ that forces writers to avoid overly simple references, and one sentence under $N_c < 10$ that forces writers to produce such simple references as *I guess so* or *That sounds good*.

[1] https://www.google.co.jp/trends/topcharts#date=2012

Input sentence	References	Sources	Winning rates
I don't like Disneyland when it's very crowded... そして、ディズニーランドの大混雑も苦手です …。	It's so insane when everyone starts dashing at the same time as the gates open. 開門と同時に皆が走り出す光景って、何度見てもぞっとしますよね。	Human 0%	0.96
	Oh, it was really bad.. あらら、それは大変でしたね。	Human 30%	0.43
	I'm surprised that anyone would have such a pet. 飼っている人がいると聞いてびっくりです。	Human 60%	0.01
	Yeah, I agree! あーあたしも！	IR-status	0.81
	It's so crowded! 大混雑だな！	IR-response	0.29
	Yes, I go to Disneyland over ten times a year. はい、ディズニーには年に 10 回は行きます。	Rule	0.20
I just checked my iTunes, and I know all of my songs are from animated movies, games, vocaloids, voice actors and audio dramas of comics. iTunes に入ってるの確認したらアニソンとゲーソンとボカロと声優さんとドラマ CD だらけだった	You must really like anime and games! アニメとかゲームが好きなんだね！	Human 0%	0.95
	Don't you listen to rocks or western music? ロックや洋楽は聞かないんですか？	Human 30%	0.88
	Whet is your favorite year of it? 一番あたりだった年はいつですか？	Human 60%	0.15
	That's normal for myself. いつものわたしである	IR-status	0.26
	Vocaloids and anime songs lol. ボカロとアニソンだね w	IR-response	0.43
	What anime songs do you like? アニソンは何が好きなんですか？	Rule	0.43

Table 2 Examples of input sentences, reference sentences and their winning rates.

The following are the details of the masked input sentences. For all input sentences, six writers created references for them without masks, two writers created references for 30% masked input sentences, and two writers created references for 60% masked sentences. We randomly assigned the input sentences to writers who imagined the masked terms and created references. They wrote 70 references for each input sentence: 42 sentences without masks, 14 with 30% masked, and 14 with 60% masked (Table 1).

In addition to the manually created references, we gathered 30 possibly negative reference sentences that were generated by the following two retrieval-based generation methods, *IR-status* and *IR-response* [3], and one rule-based generation method, *Rule* [1]. *IR-status* retrieves reply posts whose associated source posts most closely resemble the input user utterances. The *IR-response* approach is similar to the *IR-status*, but it retrieves the reply posts that most closely resemble the input user utterances. *Rule* represents a rule-based conversational system that uses 149,300 rules (pattern-response pairs) written in AIML [8] and retrieves responses whose associated patterns have the highest word-based cosine similarity to the input sentence. Each method generated ten reference sentences for each input sentence.

After the reference collection, two human evaluators annotated the winner of each reference pair in terms of its *naturalness as a response*. With 100 references for each input sentence, they annotated 4,950 pairs for each input sentence. Table 2 shows examples of the input sentences and the references with their winning rates.

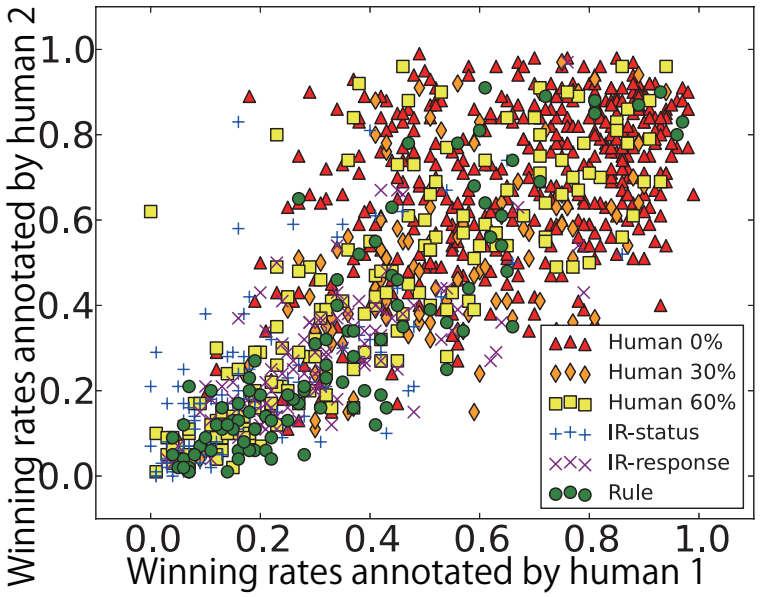

Fig. 1 Distribution of annotated winning rates between annotators

3.1.3 Estimation procedure

We compared the three methods described in Section 2.3 with smoothed BLEU that calculates BLEU over multi-references (m-BLEU)[2] and ΔBLEU [5]. All the estimations were conducted through the leave-one-out method; i.e., the methods estimated the evaluation scores for each reference sentence using the other 99 references. The parameters of the methods are experimentally determined. We used 3 for the M of AM (Average of metrics) and WS (Weighted scores), SVR with RBF-kernel, and C=5. Similarity metrics s used in AM, WS, and Regression are either sentence-BLEU (BLEU), RIBES [9], or Word Error Rate (WER). Here, WER, which is calculated as normalized Levenshtein distance NL to a reference sentence, is converted to a similarity with either WER=1−NL (ranges from 0 to 1) or WER=1−2NL (-1 to 1).

3.2 Analysis of annotated evaluations

Before the experiments, we performed a brief analysis of the manually annotated evaluation scores (winning rates). Fig. 1 shows their distribution between the annotators. They are broadly distributed along the whole range of 0-1. The manually

[2] We used NIST geometric sequence smoothing, which is implemented in nltk (Method 3).

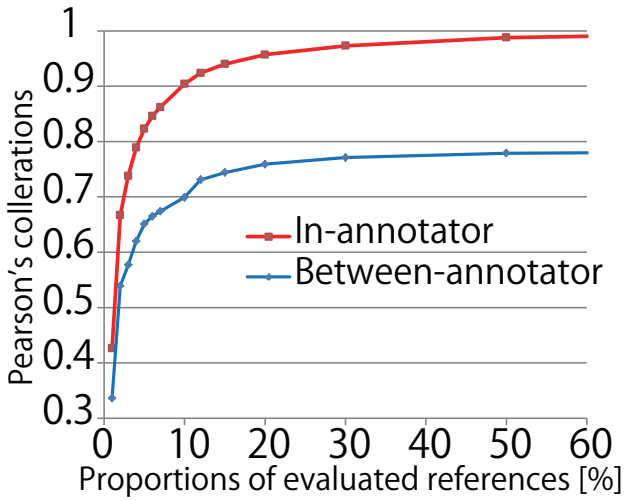

Fig. 2 Annotated pairwise propotions vs. correlations

created references (red triangles, orange diamonds, and yellow squares) were evaluated as more natural than the system-generated references. Comparing the system-generated references, those generated from the retrieval-based methods (*IR-status*: blue crosses, *IR-responses*: purple x marks) are gathered in the low or middle winning rates, and those generated from *Rule* (green circles) are distributed along the whole range. This shows that *Rule* generated references with the same appropriateness as the manually created ones when the rules correctly matched the input sentences. Pearson's correlation coefficient between the human evaluators was 0.783. Fig. 1 also shows that the references with low winning rates show stronger correlations since the points of lower left corner gather at y = x. This result indicates that the negative input-response pairs are consistent between the evaluators, but the positive pairs are somewhat different probably because negative ones can be checked with violation of some criteria such as Grice's maxims [10].

Figure 2 shows the variation of the pearson's correlation in- and between-evaluators over the rate of the evaluated pairs. We obtain the winning rates from partially sampled pairwise evaluations. The increase of the coefficients become slow around 12% of the evaluation rates. With our 100 references, 600 pairwise evaluations are enough to obtain the winning rates with high correlation coefficients ($r = 0.924$ in in-annotator condition and $r = 0.731$ in between-annotator condition) with the true rates.

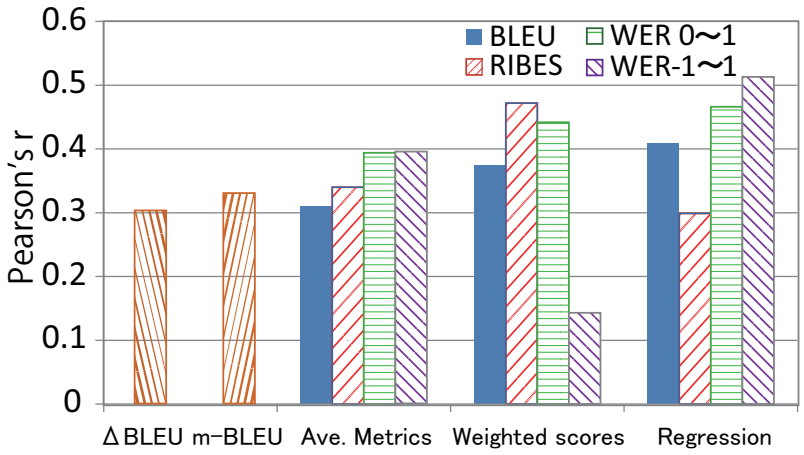

Fig. 3 Correlation between annotated (human 2) and estimated scores

3.3 Results

Figure 3 shows the correlation coefficients of the combination of the sentence similarity metrics and the proposed methods. SVR with WER (using the range from -1 to 1) shows the highest correlation ($r = 0.514$). Among the AM (Average of metrics) methods that leverage only the positive references, WER (-1 to 1) shows the highest correlation but still has lower correlations ($r = 0.399$) than those that leveraged the negative references. We calculated these scores using human 2 annotations, but it does not differ from the scores using human 1 annotations.

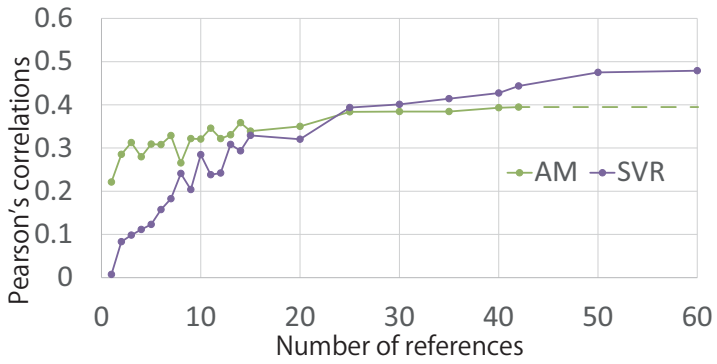

Fig. 4 Correlations over number of references

Figure 4 shows the relations between the number of references and the correlations of SVR with WER and AM with WER. With fewer references, AM shows

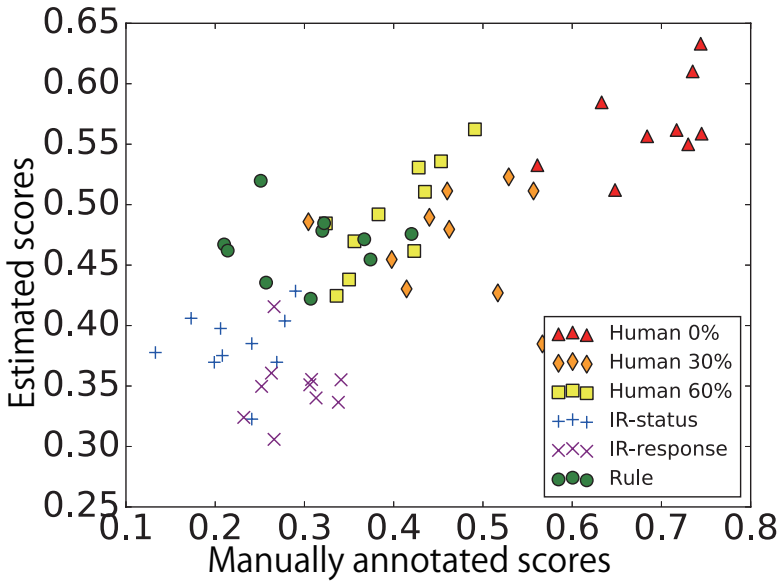

Fig. 5 System-wise comparison of annotated and estimated evaluation scores

higher correlations than SVR, because it requires training samples for accurate estimations, while AM can output reasonable estimations even with just one reference. The SVR performance becomes higher than AM with over 25 references and continues to improve. This indicates that both of our regression-based approach and the large size of references are keys to estimate the scores with high correlation.

Figure 5 shows the system-wise evaluation scores between manual annotation and SVR estimation. Each point is calculated as the means of ten scores; each score is sampled from the estimated scores of an input-reference pair whose references are associated with certain generation methods (e.g., human 30% mask or *Rule*). The scores are highly correlated with Pearson's $r = 0.772$. Figure 5 illustrates that the references generated from human 60% mask, *Rule*, and *IR-status* are estimated with higher scores than the manual scores. This is because most of the low-evaluated references of human 60% mask and *Rule* have correct grammar but wrong contents and are barely distinguished with WER that only considers edit counts. *IR-status* has many expressions that did not appear in other references, such as *lol* (*www* in Japanese) and emoticons like *:-)*. The differences between these expressions and the references are difficult to evaluate with WER and BLEU because they depend on word matching. This problem may be solved using character N-grams and the proportions of the character types as regression features.

4 Conclusion

We proposed a regression-based evaluation method for chat-oriented dialogue systems that appropriately leverages many positive and negative references. The sentence-wise correlation coefficient between our proposed and human annotated scores reached 0.514 and the system-wise correlations were 0.772. These scores are significantly higher than the previous methods such as delta-BLEU that define evaluation scores with sentences similarities between system output and reference sentences. Our results indicate that both of our regression-based approach and the large size of references are keys to estimate the scores with the high correlation. The limitation of our work are the small number of inputs and the huge cost of winning rates. We are planning to large-scale input-reference pairs with Likert scale evaluations to examine the effectiveness of our approach and the differences between winning rates and Likert scales.

References

1. Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing. In: Proceedings of the 25th International Conference on Computational Linguistics. pp. 928–939 (2014)
2. Sugiyama, H., Meguro, T., Higashinaka, R., Minami, Y.: Open-domain Utterance Generation Using Phrase Pairs based on Dependency Relations. In: Proceedings of Spoken Language Technology Workshop. pp. 60–65 (2014)
3. Ritter, A., Cherry, C., Dolan, W.: Data-Driven Response Generation in Social Media. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 583–593 (2011)
4. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on Association for Computational Linguistics. pp. 311–318 (2002)
5. Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., Dolan, B.: Delta-BLEU : A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. pp. 445–450 (2015)
6. Sculley, D.: Large Scale Learning to Rank. In: Proceedings of NIPS 2009 Workshop on Advances in Ranking. pp. 1–6 (2009)
7. Smola, A.J., Sch, B., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing 14(3), 199–222 (2004)
8. Wallace, R.S.: The Anatomy of A.L.I.C.E. ALICE Artificial Intelligence Foundation, Inc. (2004)
9. Isozaki, H., Hirao, T., Duh, K., Sudoh, K., Tsukada, H.: Automatic Evaluation of Translation Quality for Distant Language Pairs. In: Proceedings of the conference on Empirical Methods on Natural Language Processing. pp. 944–952 (2010)
10. Grice, H.P.: Logic and Conversation. In: Syntax and semantics. 3: Speech acts, pp. 41–58 (1975)