Classification-and-Ranking Architecture for Response Generation based on Intentions

Aida Mustapha, Md. Nasir Sulaiman, Ramlan Mahmod and Hasan Selamat

University Putra Malaysia, Serdang, Selangor Darul Ehsan, Malaysia

Summary
Grammar-based natural language generation is lacking robustness in implementation because it is virtually incapable for learning. Statistical generation through language models is expensive due to overgeneration and its bias to short strings. Because dialogue utterances render intentions, learning model for the response generation systems should consider all utterances as equally good regardless of length or grammar. An intention-based architecture has been developed to generate response utterances in dialogue systems. This architecture is called classification-and-raking. In this architecture, response is deliberately chosen from dialogue corpus rather than wholly generated, such that it allows short ungrammatical utterances as long as they satisfy the intended meaning of input utterance. The proposed architecture is tested on 64 mixed-initiative, transaction dialogue corpus in theater domain. The results from the comparative experiment show 91.2% recognition accuracy in classification-and-ranking as opposed to an average of 68.6% accuracy in overgeneration-and-ranking.

Keywords:
Intentions, Speech Acts, Dialogue System, Natural Language Generation, Classification-and-Ranking.

1. Introduction

In human-human conversation, dialogue is mutually structured and timely negotiated between dialogue participants. Speakers take turns when they interact, interrupt each other, but their speeches seldom overlap. Similarly, human-machine conversation using dialogue systems exhibits comparable qualities. A response generation system is the natural language generation component in dialogue systems, which is responsible for providing dialogue responses as part of interactive human-machine conversation.

The high degree requirement of linguistic input specifications in grammar-based natural language generation is the classic problem of knowledge engineering bottleneck. Statistical generation through language models, although robust, is expensive because alternative realizations and their probabilities have to be calculated individually. Language models also have built-in bias to short strings because likelihood of a string of words is determined by joint probability of words. This is not desirable for generation in dialogues because utterances render intentions; hence all realizations should be treated as equally good regardless of length, in fact, regardless of grammar. The main focus of this paper is to propose a new architecture for response generations based on intentions.

The remainder of this paper will be organized as follows. Section 2 will begin with discussion of related works in natural language generation. Section 3 will introduce intentions; the basic building block to our response generation architecture. Section 4 will present the two-staged classification-and-ranking architecture while Section 5 will present validation experiments to compare the proposed architecture with the existing overgeneration-and-ranking architecture. Finally, in section 6 we will draw some concluding remarks.

2. Related Works

Existing architectures for natural language generation in dialogue systems mainly concern with generation of words into sentences, either by means of grammar or some statistical distribution. Grammar-based approach requires specification of fine-grained meaning representations as input to guide the generation process. Because this process expects a large number of explicit features, it leads to knowledge acquisition bottleneck in both constructing and maintaining the hand-crafted rule systems [1]. To alleviate the knowledge engineering load in grammar-based approach, the statistical approach of overgeneration-and-ranking architecture [1-6] provides the necessary linguistic decisions through statistical models trained on corpus to furnish semantically related utterances.

Overgeneration-and-ranking architecture combines rule-based overgeneration with ranking based on statistical or language models. The principle objective is to help

Manuscript received December 5, 2008
Manuscript revised December 20, 2008

reducing the amount of syntactic knowledge to be hand-coded manually as required by grammar-based approach. Nonetheless, the main limitation to overgeneration-and-ranking architecture is that, it is computationally very expensive because the need to overgenerate thousands of utterance candidates, either through simple grammar rules or language models like *n*-grams [3]. In addition, ranking through language models are also biased towards shorter strings because the likelihood of a string of words is determined by the joint probability of the words [7].

Overgeneration-and-ranking also work well in written language where sentence is the basic unit. However, in spoken language where utterance is the basic unit, the disadvantage becomes critical as spoken language also render intentions, hence short strings may be of equivalent impact. This approach is clearly not necessary for generation of dialogue utterances that are not necessarily realized as complete sentences, and ranking must also be able to treat all candidates as equally good realizations regardless of length, in fact, regardless of grammatical formation. Because of this, even though the resulting utterances are inarguably sophisticated, the impact may not be as forceful.

We believe that the design for response generation requires more than grammar rules or some statistical distributions, but more intuitive in the sense that learning of system's responses robustly satisfies the intention of input utterance. This means that response generation must be constrained by the content of intentions, rather than the lexicons and grammar.

3. Intentions

At the most abstract level, we develop the classification-and-ranking architecture for response generation in dialogue systems on the basis of speech actions by Austin [8]. He introduces the idea of "speech action", which gives account for the functional meaning of an utterance, rather than only the truth-conditional of utterance interpretation. The idea is that an utterance is essentially a request for action, not a request for information. The theory also focuses on communicative acts performed through speech utterances, in a tripartite classification called "speech acts". Speech acts are classified into locution (the actual words uttered to deliver the utterance), illocution (the force or intention behind the words in the utterance), and perlocution (the effect of the intention on the hearer).

Searle [9] continues to polish communicative acts at the illocutionary level through Speech Acts Theory, with fine-grained characterization of illocutionary acts into categories like assertive acts, directive acts, permissive acts, and prohibitive acts. Ultimately, Searle believes that the illocutionary force of sentences, or intentions, is what determines the semantics of language. However, both [8, 9] only address the effect of speech acts with regards to the context in isolated utterance, but not the consequence of speech acts produced by a series of utterances in a conversation. To account for speech acts within a stretch of dialogue, subsequent dialogue theories place speech acts in a bigger framework. Speech acts within the context of grounding results in a more sophisticated model of joint action in dialogue, known as "dialogue acts" [11].

Nonetheless, our intention-based architecture capitalizes on intentions beyond the context of dialogues acts to integrate speech acts with other levels of acts. We follow Conversation Acts Theory [10] to include turn-taking acts and argumentation acts as basis of classification. Turn-taking acts resolve turns that are shared between two speakers. Argumentation acts are high-level discourse acts that give shape to the entire discourse, normally referred as discourse model, discourse plan, or dialogue strategies. Argumentations acts are made by combination of smaller acts that define a task (i.e., gaining information), hence it is highly specific to domain of the systems.

4. Classification-and-Ranking

We define classification-and-ranking response generation as a deliberate process of classification and ranking response utterance from the dialogue corpus; constrained by intentions of previously contributed utterance. In the essence, response is deliberately chosen from dialogue corpus rather than wholly generated, such that it allows short ungrammatical utterances as long as they satisfy the intended meaning of input utterance. This means that the generation system learns to manage its own response strategies based on corpus.

Overgeneration-and-ranking architecture employs two basic components as illustrated in Figure 1. The first component is a classifier that classifies user input utterances into response classes based on their contextual, pragmatic interpretations. The second component is a ranker that scores the candidate response utterances according to semantic content relevant to the input utterance. The mechanic is to find the response class where the possible responses reside in and next to access the probabilities of response candidates in that particular response class, which satisfies the intentions of user input utterance.

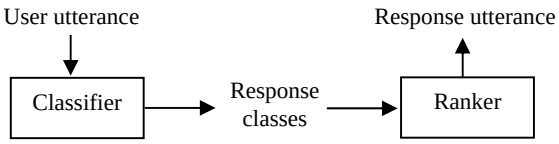

Fig. 1 The two-staged classification-and-ranking architecture.

In this architecture, the basic assumption is that for every pair of utterance exchange in the dialogue corpus, the first pair from the user creates certain expectations that can constrain the possibilities of the second pair from the response. This means that during learning, each user utterance must be represented in some pragmatic knowledge form so it can be unique to its counterpart response from dialogue corpus. The pragmatic features to assist the classifier rely on dialogue acts as well as turn-taking acts, grounding acts, and argumentations acts based on Conversational Acts Theory [10].

4.1 Meaning representations

To support the architecture, we advocate for unified meaning representation in utterances so the response generator is able to acquire input directly from the natural language understanding (NLU) component. The input is a set of dialogue acts that consists of two elements; the communicative functions and the propositional content as formally described by [11, 12]. Communicative functions represent illocutionary force (intentions) in the utterance, while semantic content corresponds to contribution of utterance to the context so far. The elements are represented as an *input frame* in the form of attribute-value pairs. Figure 2 illustrates an example of input frame *u* of user input utterance *U*.

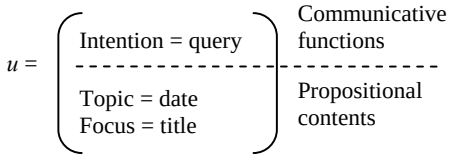

Fig. 2 Input frame for user input utterance.

While communicative functions are the building blocks for the classification stage, the application of propositional contents is leveraged to both stages; classification and ranking. During classification, we require the topic of utterance to aid in modeling the pragmatic representations of the utterance, which are turn-taking and argumentations acts. On the other hand, during ranking, we adopt focus of attention in the particular utterance as key constraint to weigh response utterances in the respective response class.

The extraction strategy for topic and focus is based on analysis of Information Structure Theory [13] with respect to the informational point of view. According to this theory, topic is essentially what an utterance is about; it contains known information based on context of conversation. Focus, on the other hand, is new information that requires special attention. Halliday [13] adopts topic articulation as the first element in the utterances, which is essentially the subject, while focus is the object. Hence, topic and focus depends on the mood of utterances whether it is an assertive, imperative, or interrogative. In each mood, subject and object of utterance occupy different position depending on the structure.

4.2 Response classification

The main task of classification in the classification-and-ranking architecture is to determine which set of response class that a user utterance belongs to. A response class contains all response utterances that pairs to a particular input utterance based on dialogue corpus. In terms of probability, we would like to identify a response class in which the counterpart response utterance resides in, such that P (*response class* | *user utterance*) is highest. Feature variables are represented by pragmatic properties of input utterance that we wish to classify. Equation 1 shows the equation for picking the best response class. Next, by using Bayes rule, we recompose the equation for our response class probability into three factors as the final decision rule as shown in equation 2.

$$\hat{rc} = \arg\max_{rc \in R} P(rc \mid U) \tag{1}$$

$$\hat{rc} = \arg\max_{rc \in R} \frac{P(U \mid rc)P(rc)}{P(U)} \tag{2}$$

The output of classification stage is a response class coherent to the input utterance as learned from the pair of user-system utterances in the dialogue corpus. Classification of user utterances into response classes is imperative in order to delimit the searching space for final response utterance during ranking. Table 1 shows the features employed during classification of user utterance into the corresponding response class. The features are driven by Conversation Acts Theory [10] with respect to our intentional point architecture. These features are employed during classification to constrain response utterance according to their contextual contributions, therefore, guiding the ranking process to find one single response utterance that is most relevant to the input utterance.

Table 1 Features for classification of user utterance.

No	*Features*	*Desciptions*
1	Forward-looking functions (*FLF*)	Speech act for user utterance
2	Backward-looking functions (*BLF*)	Grounding act for user utterance
3	Context (*CX*)	Global topic of user utterance
4	Topic (*T*)	Topic of user utterance
5	Mood (*M*)	Mood of user utterance i.e., declarative, interrogative or imperative.
6	Control (*C*)	Control holder at the point of user utterance
7	Role (*R*)	Role of the user
8	Turn (*TU*)	Turn-taking act for user utterance
9	Argumentation (*ARG*)	Argumentation act of user utterance
10	Response Class (*RC*)	Response class that is tagged to user utterance

4.3 Ranking utterances

Given the response class provided by classification stage, ranking assigns probability score to each response utterance in the particular response class. The response class *RC* holds possible response utterances $\{r_1 r_2 ... r_R\}$ from the set of responses *R*. The goal of the ranking is to output a single response utterance $r \in \{r_1 r_2 ... r_R\}$ in respond to the user by choosing the response utterance that yields highest probability.

Ranking assumes two kind of input; the response class and focus of user utterance as provided in the input frame. According to Grosz [14], while topic anchored to the common about of the utterance, focus delivers the very information that the utterance is structured to convey. Therefore, focus of attention tracks the entities as the dialogue progresses. Table 2 shows the features employed during ranking to distinguish response utterances as unique from one to another.

Table 2 Features for ranking of response utterance.

No	*Features*	*Desciptions*
1	FLF	Speech act for response utterance
2	BLF	Grounding act for response utterance
3	Topic	Topic of conversation in response utterance
4	Focus	Focus of attention in response utterance
5	uFocus	Focus of attention in user utterance
6	DA	Domain attributes in response utterance

Apart from communicative functions (Feature 1-2) and propositional contents (Feature 3-5), the decision to assign the probability scores is also guided by the informativeness measure in response utterances. Because dialogue corpus is domain-specific, response utterances must be semantically represented using some ontology general enough for future reuse in another domain. In our architecture, we access the information value through abstraction of semantics encoded in each response utterances into domain-specific attributes as shown as Feature 6 in Table 2.

The probability model is defined over *R* x *S*, where *R* is the set of possible response utterances $\{r_1 r_2 \ldots r_R\}$ and *S* is the set of corresponding semantic features to each response utterances. The set *S* consists of both local and global knowledge for the response database *R*. Local knowledge are features extracted from response utterances in training corpus, represented by the presence or absence of domain attributes, for example *title*, *genre*, or *date*. Global knowledge is supplied by focus of attention in user utterance. Using both local and global features to model the probability distribution, *M* feature functions $f_m(r, \{r_1 r_2 \ldots r_R\}, s)$ were defined for each evidence in the training data where $r \in R$, $s \in S$ and $m = 1,\ldots,M$. The probability model of response utterance *r* is conditioned to features *s*, where λ_m are the weights associated with each feature *m*. The decision rule is shown in the following equation 3.

$$\hat{r} = \arg\max_{r \in R} \left[p(r \mid \{r_1 r_2 ... r_R\}, s) \right] \tag{3}$$

where

$$p(r \mid \{r_1 r_2 ... r_R\}, s) = \frac{1}{Z(s)} \exp\left[\sum_{m=1}^{M} \lambda_m f_m (r, \{r_1 r_2 ... r_R\}, s) \right]$$

and

$$Z(s) = \sum_{r'} \exp\left[\sum_{m=1}^{M} \lambda_m f_m (r', \{r_1 r_2 ... r_R\}, s) \right]$$

5. Validation Experiments

Performance of the proposed classification-and-ranking architecture presented in this paper is evaluated with respect to ranking accuracy if the top-ranked response utterance returned by the ranker is correct or otherwise according to user utterance in the test set. The experiment is repeated for overgeneration-and-ranking techniques, which are language models (LM), Maximum Entropy with language model (ME-LM), and instance-based learning (IBL). This section evaluates and compares the relative performance of existing overgeneration-and-ranking and the proposed classification-and-ranking architecture. We

trained and tested our architecture separately all 15 response classes. Our evaluation metric is based on recognition accuracy of the response utterances with regard to a user input utterance. The baseline accuracy that would be expected if we were to randomly pick a response utterance is 21.8%.

5.1 SCHISMA Corpus

SCHISMA is an acronym derived from the Dutch *SCHouwburg Informatie Systeem*, a theater information and ticket reservation system developed in the University of Twente [15]. This system enable users to reserve a particular show from a wide range of available options but both user and system must collaborate to achieve an agreement to several issues like the ticket price, the seating arrangement or the discount availability. The corpus has been annotated with dialogue acts using Dialogue Act Markup in Several Layers (DAMSL) by Keizer [16]. Figure 3 shows an extract of SCHISMA dialogues.

U:	What will be on in the Theatre next week (19 March)?	[1]
S:	There is no show on that date.	[2]
U:	And on 18 March?	[3]
S:	In the period 18 March 1994 until 20 March 1994 you can go to Deelder Denkt and Indonesian Tales.	[4]
U:	At what time does Deelder start?	[5]
S:	The show starts at 20:00.	[6]
U:	How much does it cost	[7]
U:	and are there still places?	[8]
S:	Do you have a reduction card?	[9]
U:	No	[10]
S:	The price for the show Deelder Denkt is f26,00.	[11]
S:	And there are still 82 places free.	[12]

Fig. 3 SCHISMA dialogue excerpt.

The entire corpus is made of 64 dialogues, whereby 920 utterances are from a human user and 1,127 utterances from the system. Response utterances are then collapsed according to turn in order to assign one unique response to each user input utterance. 920 response utterances are then classified into 15 response classes based on topic of the utterances, which are *title*, *genre*, *artist*, *time*, *date*, *review*, *person*, *reserve*, *ticket*, *cost*, *availability*, *reduction*, *seat*, *theater*, and *other*.

5.2 Results

We evaluate the final output of intention-based response generation based on relevance by claiming that a response is relevant only when it satisfies the intention of the preceding utterance. Evaluation is performed by judging the final response utterance returned as the top-ranked response by the system is an equivalent or otherwise when compared to the actual response provided by the dialogue corpus. Note that we use the term "equivalent" rather than "identical" based on the ground that there could be more than one response utterance that conveys the same semantic and pragmatic interpretation although in different form of surface structures. An equivalent response also reflects that the response is coherent to the dialogue context and relevant to the preceding input utterance. This is different from overgeneration-and-ranking architecture that judges the response utterances based on 'fluency' of output utterance compared to word pairs from the dialogue corpus.

The results from the comparative experiment show 91.2% accuracy in the proposed intention-based ranking as opposed to an average of 68.6% accuracy in overgeneration-and-ranking. Distribution of accuracy is shown in table 3.

Table 3 Accuracy percentages overgeneration-and-ranking compared with the proposed classification-and-ranking.

NO	*RC*	*LM*	*ME-LM*	*IBL*	*PRO-POSED*
1	title	79.2	59.6	44.2	91.3
2	genre	95.9	96.4	67.9	89.3
3	artist	60.0	76.2	40.0	90.5
4	time	78.8	68.8	56.3	90.6
5	date	83.1	71.1	54.3	93.3
6	review	41.2	41.1	16.1	89.3
7	person	69.4	76.7	56.7	96.7
8	reserve	92.4	52.7	38.8	79.3
9	ticket	83.3	87.7	65.4	97.5
10	cost	95.4	64.2	52.6	90.6
11	avail	46.2	78.6	40.0	92.9
12	reduc	83.2	93.2	72.0	93.2
13	seat	84.4	81.9	60.6	93.6
14	theater	61.5	91.7	68.0	100
15	other	76.5	95.1	89.5	80.3
	Avg	*75.4*	*75.6*	*54.8*	*91.2*

Recall that classification-and-ranking architecture advocates for intentionality and informativeness in each stage, respectively. Learning response utterances enable the generation system to replace at least some part of the knowledge engineering effort to construct the grammar rules. This is because linguistic decisions are statistically learned, hence the system does not require fine-grained input specifications. Also, the significant advantage of this architecture is that the same knowledge representation may be used to support both domain and linguistic

knowledge [17]. This is achieved through similar representations of dialogue acts in natural language understanding (NLU) module and the response generation module. In overgeneration-and-ranking, this is not necessarily being the case.

6. Conclusions

In this paper, we provide an account of a computational mechanism to characterize response utterances provided by a dialogue corpus and to reuse them in responding to new input. Our intention-based response generation employs two basic components. The first component is a classifier that classifies response utterances into response classes based on pragmatic interpretation of input utterance. The second component is a ranker that scores the response utterances in a particular response class based on informativeness of the response utterance. This is an alternative design of NLG component in dialogue systems to cater for shortcomings in grammar-based and statistical natural language generation.

In conclusion, because dialogues are intention-driven, the major concern in response generation is therefore the coherence of the entire dialogue. We argue that a response is relevant when it satisfies the intention of the preceding utterance. Hence, response realization must be based on intentions of the input utterance, rather than its syntactic form.

References

[1] Varges, S. Instance-based Natural Language Generation. Ph.D. Thesis, School of Informatics, University of Edinburg. 2003.
[2] Knight, K. and V. Hatzivassiloglou. Two-level, Many-path Generation. In Proceedings of 33rd Annual Meeting of the Association for Computational Linguistics. 1995.
[3] Langkilde, I. and K. Knight. Generation that Exploits Corpus-based Statistical Knowledge. In Proceedings of the 36th Annual Meeting on Association for Computational Linguistics, Montreal, Quebec, Canada. 1998.
[4] Oberlander, J. and C. Brew. Stochastic Text Generation. Philosophical Transactions of the Royal Society of London. 2000, 358(A):1371–1385.
[5] Langkilde, I. Forest-based Statistical Sentence Generation. In Proceedings of the North American Meeting of the Association for Computational Lingustics (COLING-00). 2000.
[6] Bangalow, S. and O. Rambow. Exploiting A Probabilistic Hierarchical Model for Generation. In Proceedings of the 18th International Conference on Computational Linguistics, Saarbrucken, Germany. 2000.
[7] Belz, A. Statistical Generation: Three Methods Compared and Evaluated. In Proceedings of the 10th European Workshop on Natural Language Generation (ENLG 05). 2005.
[8] Austin, J. How to do Things with Words. Oxford: Clarendon. 1962.
[9] Searle, J. Speech Acts. Cambridge University Press. 1969.
[10] Traum, D. and E. Hinkelman. Conversation Acts in Task-oriented Spoken Dialogue. Computational Intelligence. 1992, 8(3):575–599.
[11] Bunt, H. Dialogue Pragmatics and Context Specification. In H. Bunt and W. Black (Eds.), Abduction, Belief and Context in Dialogue. Studies in Computational Pragmatic, Amsterdam, Benjamins. 2000, pages 81–150.
[12] Hulstijn, J. Dialogue Models for Inquiry and Transaction. Ph.D. Thesis, University of Twente, Netherlands. 2000.
[13] Halliday, M. Notes on Transitivity and Theme in English. Journal of Linguistics 3, 1967.
[14] Grosz, B. The Representation and Use of Focus in a System Understanding Dialogs. In IJCAI-77, Morgan Kauffman. 1977, pages 67–76.
[15] Hoeven, v.d. G, Andernach, A., Burgt, v.d. S., Kruijff, G., Nijholt, A., Schaake, J., and Jong, F. SCHISMA: A Natural Language Accessible Theater Information and Booking System. In Proceedings of the 1st Internation Workshop on Applications of Natural Language to Data Bases. 1995, pages 271-285.
[16] Keizer, S. Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks. Ph.D. Thesis, University of Twente, Netherlands. 2003.
[17] Reiter, E. and C. Mellish. Using Classification to Generate Text. In Proceedings of the 9th COLING. 1992.

Aida Mustapha received the B.Sc. degree in Computer Science from Michigan Technological University and the M.IT. degree in Computer Science from Universiti Kebangsaan Malaysia in 1998 and 2004, respectively. She recently received her Ph.D. in Artificial Intelligence focusing on dialogue systems. She is currently an active researcher in the area of Computational Linguistics and Conversational Agents.