Intention-based Ranking for Surface Realization in Dialogue Systems

Aida Mustapha aida@fsktm.upm.edu.my
Faculty of Computer Science and Information Technology
University Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Md. Nasir Sulaiman nasir@fsktm.upm.edu.my
Faculty of Computer Science and Information Technology
University Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Ramlan Mahmod ramlan@fsktm.upm.edu.my
Faculty of Computer Science and Information Technology
University Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Mohd. Hasan Selamat hasan@fsktm.upm.edu.my
Faculty of Computer Science and Information Technology
University Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia

Abstract

A new intention-based ranking is proposed to cater for intentionality in ranking dialogue utterances, as opposed to surface-based ranking using surface linguistic features in utterances. This is because utterances may be in the form of a sentence, a phrase, or just a word; hence basis for ranking must be on assessment of intentions, regardless of length of utterance and grammar rules. Intention-based ranking model is tested and compared with surface-based models on 15 response classes in theater domain. The results from comparative experiments show consistent accuracy distribution in intention-based ranking across all response classes with average of 91%. On the contrary, ranking accuracy for surface-based ranking is not uniform across the response classes, showing the dependency on surface representation of utterances in individual response class.

Keywords: Overgeneration-and-ranking, Ranking, Classification, Intention, Dialogue systems.

1. INTRODUCTION

Statistical approaches to surface realization sidestep the linguistic decision-making process by applying statistical learning in the surface generator itself, as opposed to the deterministic knowledge-based approach. This approach is known as overgeneration-and-ranking [1], which relies on corpus to furnish semantically related sentences through surface linguistic features of sentences. The principle objective is to help reducing the amount of syntactic knowledge to be hand-coded manually as required by knowledge-based approach. The effort required to construct grammar for the overgenerator is also very minimal; enough for it to generate lattices. Due to the minimal generation technology, an additional task of ranking is necessary. The need for ranking arises to discriminate out candidate sentences that are ungrammatical, unintelligible or at least not fluent by means of language models.

Langkilde and Knight [1] and Langkilde [2, 3] focused on learning surface structure of sentences at the syntactic level, while subsequent researches [4, 5] extended learning into semantic level through incorporation of semantic information. This is essentially a mapping from semantic to syntactic. For instance, Bangalore and Rambow [4] use dependency tree labeled with extended synonyms rather than lexemes, while Varges [5] utilizes semantic mark-up in constructing its grammar base. Similarly, Ratnaparkhi [6] and Oh and Rudnicky [7] apply language models on ranking generation templates. Nonetheless, the motivation remains, which is to learn and regenerate the sentences based on surface linguistic features.

The main limitation of overgeneration-and-ranking is that, it is computationally expensive to overgenerate in setting up the band of realization candidates, either through simple grammar-rules or statistical means like *n*-grams [5]. While knowledge-based approach through grammar is not usually fast enough for use in dialogue systems [8], overgeneration is also not necessary for generation of dialogue utterances due to two main reasons. Firstly, dialogue utterances are typically short, single-sentenced, and are often incomplete. They can take form of a sentence, a phrase, or just a word. Secondly, dialogue utterance bears individual intention. Even if the surface form is grammatically incorrect, an utterance fares well as long as it satisfies the intentions of the utterance it is responding to.

Language models, although robust, also have built-in bias to produce short strings because the likelihood of a string of words is determined by the joint probability of the words [9]. This is clearly not desirable for generation of dialogue utterances because all utterance should be treated based on assessment of the intentions, regardless of length, in fact, regardless of grammar. While the output may be inarguably sophisticated, the impact may be not as forceful. We believe that ranking dialogue utterances requires more than statistical distributions of language, but more intuitive in the sense that ranking model incorporates intentionality to maintain coherence and relevance, regardless of the surface presentation.

Intention-based ranking [10] is taking pragmatic approach to assessing dialogue utterances. Different from previous ranking models that deal with language models and semantic features, intention-based ranking focuses on finding the best utterance based on the semantic and pragmatic knowledge they represent. The knowledge exists in the form of (1) semantics from user utterances and (2) intentions, semantics, and domain informativeness from response utterances. The utterance with highest probability score is said “relevant” with respect to input utterance when topic of response utterance satisfies the intention of user utterance.

The remainder of this paper is organized as follows: Section 2 will present four different ranking models; three of the models are surface-based while the last is the proposed intention-based ranking model. Section 3 will provide experimental background by introducing the corpus and dataset used during the experiment. Finally, result findings are reported and discussed in Section 4 before the paper is concluded in Section 5.

2. RANKING MODELS

Surface-based ranking under the overgeneration-and-ranking methodology involves a task to rank all sentences or utterances (called lattices) resulted from an overgeneration process that capitalizes on semantic and surface linguistic features obtained from the corpus. The goal is to find the highest probability utterance ranked as output of the process. Similarly, the goal of intention-based ranking is also to find an utterance with the highest probability as the output. Nonetheless, while surface-based ranking may consider hundreds or thousands of lattices at one time, intention-based ranking only consider utterances in specific, individual response class, resulted from the classification process under the classification-and-ranking methodology.

This section presents decision rules for all surface-based ranking models that we consider; which are *n*-grams language model, maximum entropy with language model, and instance-based learning model. At the end of the section is the decision rule for the proposed intention-based ranking model that capitalizes on intentions rather than surface features.

2.1 *N*-grams Language Model

A language model is a statistical model of sequence of words, whereby probability of a word is predicted using the previous *n*-1 words. Following *n*-gram ranking [1, 11, 12], response utterances are trained by a trigram model through counting and normalizing words inside the utterances. Consider the following response utterance:

r = "*Yes there are still 826 seats available.*"

A trigram generation model for response utterance *r* will record the pair (still, 826) or (826, seats) and the triple (still, 826, seats) or (826, seats, available). The model will then estimate the probability for *P(r)*, which is the estimated probability for response *r* based upon some count *C*. Therefore, to estimate the probability that "seats" appears after "826", the model divides the count of the pair (826, seats) by the triple (826, seats, available). This ratio is known as relative frequency or maximum likelihood estimation.

Equation 1 shows estimation of probabilities based on relative frequency of words inside response utterance *r* and the final probability of each response utterance, where *n* is total number of running words in the utterance and w_1^n is *n*-gram of $w_1 \ldots w_n$ instances in training set.

$$P(w_i \mid w_{i-2}, w_{i-1}) = \frac{C(w_{i-2} w_i)}{C(w_{i-2} w_{i-1})} \tag{1}$$

$$P(w_1^n) = \prod_{i=1}^{n} P(w_i \mid w_{i-2}, w_{i-1})$$

Based on the equation, response utterances are ranked using the negative log probabilities with respect to the language model. Back-off smoothing was applied for unobserved *n*-grams (i.e., *n*-grams that do not exist in training set), which is bigram in case of zero-probability trigram. In case of our ranking experiment, we employed features extracted by a trigram language model.

2.2 Maximum Entropy with Language Model

Similar to language models of *n*-gram, implementation of Maximum Entropy (ME) ranking [6, 12] is also surface-based, which means they rely on surface features like frequencies of *n*-grams. Nonetheless, because the ME model is trained on a corpus of existing generation templates, this provides semantic knowledge to ranking as captured by the template attributes. The basic assumption of this model is that, the best choice to express any given meaning representation (in the form of attribute-value pairs) is the word sequence with highest probability that mentions all the input attributes exactly once [6].

Feature function $f(w_i,w_{i-1},w_{i-2},attr_i)$ is constructed based on local information captured by *n*-grams and non-local information represented by the input attributes. The *n*-gram, which in this model is a bigram, enables the system to learn the word choice (lexical choice) and word order (attribute ordering) directly from the corpus. The ME probability model with bigram and domain attributes are shown in Equation 2.

$$p(w_i \mid w_{i-1}, w_{i-2}, attr_i) = \frac{\prod_{j=1}^{k} \alpha_j^{f(w_i, w_{i-1}, w_{i-2}, attr_i)}}{Z(w_i, w_{i-1}, w_{i-2}, attr_i)} \tag{2}$$

$$\text{where } Z(w_i, w_{i-1}, w_{i-2}, attr_i) = \sum_{w'} \prod_{j=1}^{k} \alpha_j^{f(w_i, w_{i-1}, w_{i-2}, attr_i)}$$

Based on the equation, ranking is performed using probability of the sequence of words $W = w_1 \ldots w_n$ given a set of attributes *A* and length of utterance *n* in the following Equation 3. To perform our ranking experiment, we abstracted out response utterances into response templates as the set of domain attributes *A*.

$$P(W \mid n, A) = \prod_{i=1}^{n} p(w_i \mid w_{i-1}, w_{i-2}, attr_i) \tag{3}$$

2.3 Instance-based Learning

Instance-based approaches are lazy, supervised learning methods that simply store the training set examples (instances) and use them directly when a new input is to be processed. At run time, the new inputs are compared to each instance in the training set (instance base). An instance-based ranker [5] scores the candidates according to their similarity to instances in the instance based taken from the training corpus. Varges [5] uses standard information retrieval techniques for representation of instances, which is *tf.idf*. The equation for *tf.idf* is represented by Equation 4, whereby $f_{i,j}$ is the term frequencies (*tf*) and $\log_2 N/n_i$ is the inverse document frequency (*idf*).

$$w_{i,j} = tf \cdot idf = f_{i,j} \times \log_2 \frac{N}{n_i} \tag{4}$$

$$\text{where } f_{i,j} = f(w_i, w_{i-1}, ..., w_{i-n})$$

For the case of our ranking experiment, *n* is the number of words in a response utterance, while *N* is the total number of response utterances in a particular class. Therefore, term frequency (*tf*) refers to individual word frequency in a particular utterance and inverse document frequency (*idf*) refers to inverse utterance frequency in a collection of response utterance *N*. After we represented all user utterances in the form of weights $w_{i,j}$, we used a simple distance measure (normalized Euclidean distance) to find the training instance closest to the given test instance, and predicts the same class as this training instance, following nearest-neighbor approach. If multiple instances include the same (smallest) distance to the test instance, the first one found is used.

Similar to the first two approaches, *n*-grams and ME augmented with *n*-grams, instance-based ranking is also surface-based, which the goal is to find the best sequence of words that forms the templates with semantic annotation tags (attributes) in place rather than the actual values for the attributes. The ranking treats the attributes just like any other words in the templates because all utterances are represented in the form of weights $w_{i,j}$.

2.4 The Proposed Intention-based Ranking

Ranking is performed on response utterances $\{r_1 r_2 ... r_R\}$ from the set of response *R*. All the response utterances are classified together based on topical contributions of each individual utterance. The goal of ranking is to output a single response utterance $r \in \{r_1 r_2 ... r_R\}$ in respond to the user; by choosing a response with the highest probability score. 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 features to each response utterances.

The set *S* consists of both local and global knowledge for each utterance in the response database *R*. Local knowledge are features extracted from response utterances in training corpus, which includes intentions (speech acts and grounding acts), semantics (topic and focus of response utterance), and domain informativeness (domain attributes i.e. *title*, *genre*, or *date*). Global knowledge is supplied by focus of attention in user utterances. Local and global variables used in intention-based ranking model are described in Table 1.

No	Feature	Desciptions
1	*rTopic*	Topic of conversation in response
2	*rFocus*	Focus of attention in response
3	*rFlf*	Speech act for response
4	*rBlf*	Grounding act for response
5	*rDa*	Domain attributes in response
6	*uFocus*	Focus of attention in user utterance

TABLE 1: Local and Global Knowledge for *R*.

Using both local and global features to model the probability distribution, each response utterance in the training data is defined in the form of *M* feature functions $f_m(r,\{r_1 r_2 \ldots r_R\}, s)$ 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* where *Z(s)* is the normalizing function as shown in Equation 6.

the normalizing function $Z(s)$ is defined in Equation

$$p(r \mid \{r_1 r_2 ... r_R\}, s) = \frac{1}{Z(s)} \exp\left[\sum_{m=1}^{M} \lambda_m f_m(r, \{r_1 r_2 ... r_R\}, s)\right] \quad (6)$$

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

Given the modeling equation, we arrive at the decision rule as shown in Equation 7.

$$\hat{r} = \arg\max_{r \in R} \left[p(r \mid \{r_1 r_2 ... r_R\}, s)\right] \quad (7)$$

3. EXPERIMENTS

The objective of this paper is to test and compare four different statistical ranking models: an *n*-gram language model, maximum entropy (ME) augmented with language model, an instance-based learning model, and the proposed intention-based model. The corpus used in the experiment is called SCHISMA, an acronym derived from the Dutch *SCHouwburg Informatie Systeem*, a theater information and ticket reservation system [13]. Figure 1 shows an extract of SCHISMA dialogues.

U:	What will be on in the theater 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]

FIGURE 1: SCHISMA Dialogue Extract.

SCHISMA is constituted by 64 text-based dialogues of varied length. In total, there are 2,047 individual utterances in 1,723 turns. 920 utterances are user contributions and 1,127 utterances are system contributions. 920 response utterances are classified into 15 response classes based on topical contributions of the corresponding response utterances by the system [10].The list of response classes is shown in Table 2.

NO	RESPONSE CLASS	NO. OF INSTANCES
1	Title	104
2	Genre	28
3	Artist	42
4	Time	32
5	Date	90
6	Review	56
7	Person	30
8	Reserve	150
9	Ticket	81
10	Cost	53
11	Avail	14
12	Reduc	73
13	Seat	94
14	Theater	12
15	Other	61
		920

TABLE 2: Distribution for Response Classes in SCHISMA Corpus.

Ranking is performed separately on response utterances (instances) in each response class as shown in Table 2. Our evaluation metric is based on recognition accuracy of the highest-ranked response utterance as compared to the dialogue corpus. In testing all surface-based and intention-based ranking models, we used the same training and testing dataset of response classes in SCHISMA. In other words, the accuracy of ranking is evaluated by checking if the response utterance returned as the top-ranked response is correct or otherwise, with respect to response utterance in the test set.

4. RESULTS AND DISCUSSION

The performance of intention-based response generation is compared with other surface-based ranking approaches that share similar spirit of overgeneration-and-ranking.

4.1 Surface-based Ranking

This section evaluates and compares the relative performance of surface-based ranking models on 15 response classes in SCHISMA corpus. Figure 2 illustrates comparison of accuracy distribution among all techniques, which are language model (LM), maximum entropy augmented with language model ME (LM), and instance-based learning (IBL). Judging from the jagged graph curve, we can see that accuracy percentage is uneven across all response classes. This is not due to the varying number of instances in each response class, but rather due to the variation of surface structure of utterances, in the sense that how unique one utterance as compared to the other.

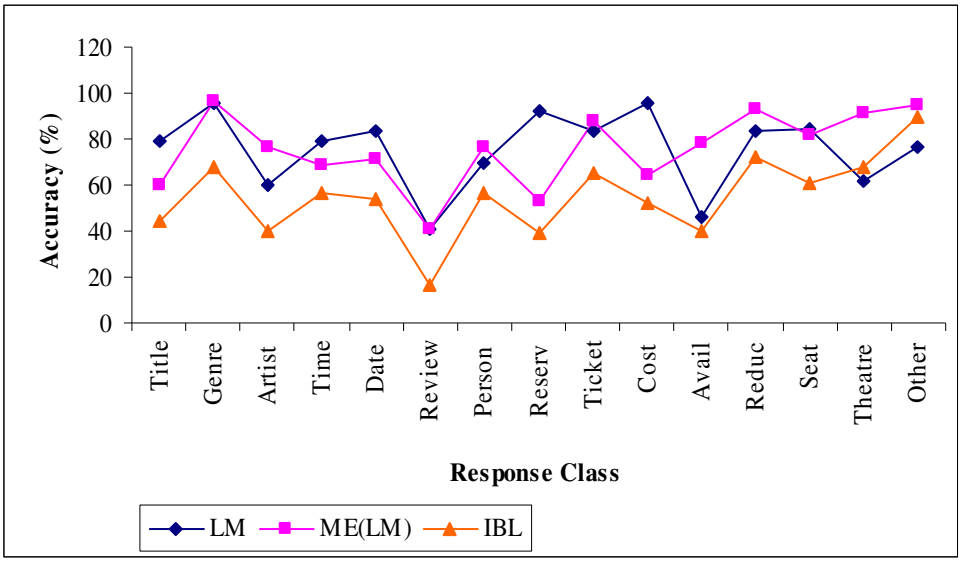

FIGURE 2: Accuracy Percentages for Surface-based Ranking.

The influence of surface forms of the utterances can be analyzed separately for all three ranking models. For trigram LM, the highest accuracy of 96% resulted from response class *genre* and the lowest accuracy of 41% from response class *review.* To study the gap between the accuracy percentages, we provide extract of instances for both response classes *genre* and *review* in Figure 3 and Figure 4 respectively.

[1] Shows fall in the following genres: Ballet, Cabaret, Dancing, family show, Musical, Music, Opera, Dramae and narrator lunch.
[2] The shows fall into the following genres: Ballet, Cabaret, Cabaret / Kleinkunst, Dance, Family Show, Youth-Theater, Musical, Music, Opera, Theater, Drama, Story-Telling and Reading. In which genre you interested?
[3] The shows fall into the following genres: Ballet, Cabaret, Dance, Family Show, Youth-Theater, Musical, Music, Opera, Theater, Drama, Story-Telling and Reading. In which genre are you interested?

FIGURE 3: Response Utterances in Response Class *genre*

[1] Show “Candide” falls in the genre musical. The famous (and traditional) love story of Candide is …
[2] Last year Han Romer and Titus Tiel Groenestege astonished the theatre world with 'Ockhams Scheermes', a show which tasted of “more”. The plans for “De Olifantsdracht” are still vague...

[3] "Der Vetter aus Dingsda" has rightfully provided Kunneke with international fame. The song "Ich bin nur ein armer Wandergesell" was and is being sung in many languages, though...

FIGURE 4: Response Utterances in Response Class *review*

Based on Figure 3 and 4, we can see that utterances in response class *genre* are short and more homogeneous in terms of the surface structures. Utterances in *review*, however, are lengthy and highly distinctive from one another. The extreme structure of response utterances in both classes shows how influential surface structures are to *n*-gram ranking accuracy.

Accordingly, the distribution accuracy for ME (LM) model is consistent with LM, except for response class *reserve*, *cost*, and *theater* that merit further explanation. Recall that for this model, *n*-gram information is in the form of domain attributes rather than the individual words in the response utterance. Ranking accuracy for both *reserve* and *cost* degrades as compared to LM because when two utterances carry the same number of domain attributes, this model is not able assign probability of the utterance any higher from the other. This can be seen from the following response utterance r_1 and r_2 that is abstracted out into *r'*.

r_1 = "You have reserved "Dat heeft zo'n jongen toch niet nodig", played by Herman Finkers on Sunday 28 May 1995. Commencement of show is 20:00. You are requested to collect these tickets minimum half an hour before commencement of the show."

r_2 = "You have reserved "De Olifantsdracht", played by Han Romer and Titus Tiel Groenestege on Sunday 22 January 1995. Commencement of show is 20:00. You are requested to collect these tickets minimum half an hour before commencement of the show."

r' = "You have reserved <title>, played by <artist> on <date>. Commencement of show is <time>. You are requested to collect these tickets minimum half an hour before commencement of the show."

Clearly, the count for domain attributes in *r'* does not help to discriminate r_1 and r_2 even though both utterances carry different semantic meaning altogether. This observation, nonetheless, does not affect response class *theater* even though the response utterances generally bear the same count of domain attributes. This is because utterances in this class carry the same semantic meaning, which is the address of the theater. Figure 5 shows excerpts of utterances in response class *theater*.

[1] Name: Theater Twent, Address: Langestraat 49, Post Code: 7511 HB, Place: Enschede, Tel.: 053-858500, Information: 053-858500.
[2] The address of this theater is: Name: Theater Twent, Address: Langestraat 49, Post Code: 7511 HB, Place: Enschede, Tel.: 053-858500, Information : 053-858500.

FIGURE 5: Response Utterances in Response Class *theater*

As for IBL ranking model, the distribution accuracy remains consistent with other models albeit the low accuracy performance as compared to LM and ME (LM). Because IBL [5] was originally implemented under the framework of text generation, this approach suffers the most due to *tf.idf* formalism used in representing weights in each response utterance. When features are represented as *tf.idf* weights, the discriminatory power for one response utterance against another degrades as *idf* assign highest weights to words that occur in few instances but still the weights will be weighed down by the *tf*.

4.2 Intention-based Ranking

The average accuracy for ranking across all response classes for SCHISMA corpus is 91.2%. To show the steady influence of intentions to ranking response utterance, the distribution of ranking accuracies is illustrated in Figure 6. Note that despite the uneven size of instances (response utterances) in every response class as shown previously in Table 2, the accuracy for all the classes is consistent from one another except for two classes, *reserve* and *other*, which merit further explanation.

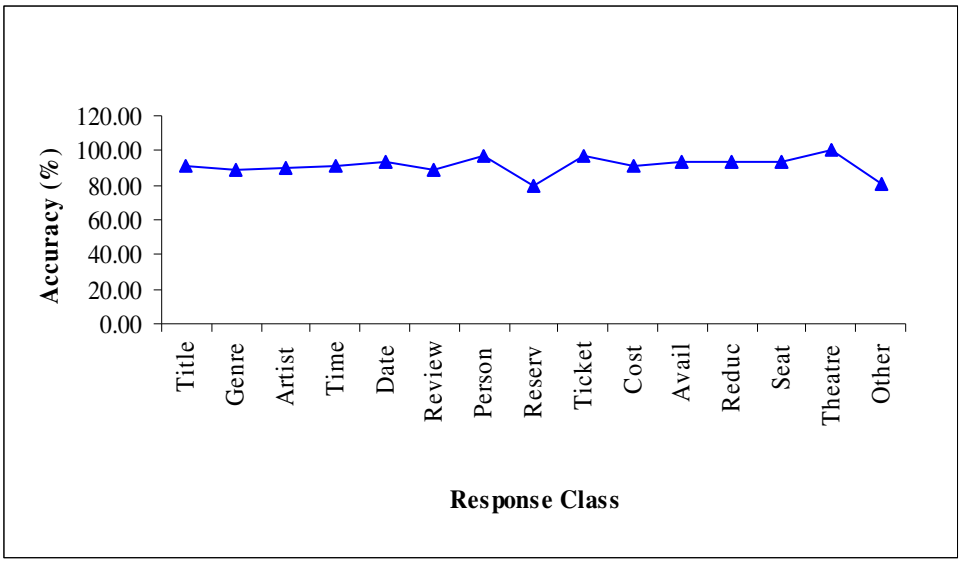

FIGURE 6: Accuracy percentages for intention-based ranking.

Figure 7 and Figure 8 show excerpt of response utterances in response class *reserve* and *other*, respectively. While response class *reserve* caters for utterances to confirm reservations, response class *other*, on the opposite, caters for utterances that do not contribute to the domain of conversation, including greeting and thanking.

[1] You have reserved “Dat heeft zo'n jongen toch niet nodig”, played by Herman Finkers on Sunday 28 May 1995. Commencement of show is 20:00. You are requested to collect these tickets minimum half an hour before commencement of the show.
[2] You have reserved “De Olifantsdracht”, played by Han Romer and Titus Tiel Groenestege on Sunday 22 January 1995. Commencement of show is 20:00. You are requested to collect these tickets minimum half an hour before commencement of the show.
[3] You have reserved 3 tickets for “Der zigeunerbaron”, played by Music Theater Prague on Saturday 11 March 1995. Commencement of show is 12.30. You are requested to collect these tickets minimum half an hour before commencement of the show.
[4] You have reserved 4 tickets for “Mevrouw Warrens Beroep”, played by The National Theater on Saturday 6 May 1995. Commencement of show is 20:00. You are requested to collect these tickets minimum half an hour before commencement of the show.

FIGURE 7: Response Utterances in Response Class *reserve*

[1] Do you want any further information?
[2] I have no information about this.
[3] Thanks for the effort and good-bye!
[4] With pleasure and Good Bye!

FIGURE 8: Response Utterances in Response Class *other*

Recall that intention-based ranking is performed based on the informativeness of the utterances; hence the interpretations of underlying semantics through domain attributes like *date*, *title*, and *other*. Albeit the variation of surface structures in both response utterances, observe that domain attributes almost all the time round down to the same attributes. For example in response class *reserve*, utterance [1] and [2] share the same set of domain attributes, which are *title*, *artist*, *date*, and *time*. Similarly, utterance [3] and [4] shares domain attributes of *ticket*, *title*, *artist*, *date*, and *time*. Assignment of domain attributes for response class *other* is no better because obviously the absence of semantics to the utterances forced the domain attribute *other*.

Since domain attributes are most likely the same in majority of the utterances, there are good chances that probabilities assigned by the ME model round down to the same figure. This leads to low accuracy results for ranking because our model is based on informativeness of utterances; hence the lack of it will average out the probability scores. At the end, one response utterances can hardly be weighed up or down from one another. Nonetheless, the impact of intention-based ranking through classification-and-ranking approach is proven to be superior to ranking based on surface features as presented by the techniques previously discussed. The ranking accuracies are consistent among the response classes, free from the influence of surface structure.

5. CONSLUSION & FUTURE WORK

Intention-based ranking differs from surface-based ranking in two major ways. Firstly, intention-based ranking apply a principled way to combine pragmatic interpretation of user utterance and the informativeness of the response utterance based on intentions, while surface-based ranking only attempts to find the best grammatical sequence of words that correspond to some meaning representations. Secondly, intention-based ranking is required to rank the utterances on the basis of relevance of a particular response utterance with regards to the input utterance. Surface-based ranking, on the other hand, is based on 'fluency' or 'completeness' of output sentences.

In the essence, as long as the technique relies on surface features [1, 5, 6], ranking accuracy is not uniform across the response classes, but rather dependent on surface representations of response utterances in individual response class. Due to this observation, in the future, we would like to investigate the performance of intention-based ranking model on other domain. Because our intention-based architecture assume the existence of dialogue act-annotated dialogue corpus based on DAMSL annotation scheme, we plan to use the MONROE corpus [15] that provides grounding and speech acts in order to run our comparative experiment.

6. REFERENCES

1. I. Langkilde 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.
2. I. Langkilde. *"Forest-based statistical sentence generation"*. In Proceedings of the North American Meeting of the Association for Computational Lingustics, 2000.
3. I. Langkilde. *"HALOGEN: A Foundation for General-Purpose Natural Language Generation: Sentence Realization Using Probabilistic Models of Language"*. Ph.D. thesis, Information Science Institute, USC School of Engineering, 2002.
4. S. Bangalore and O. Rambow. *"Exploiting a probabilistic hierarchical model for generation"*. In Proceedings of the 18th International Conference on Computational Linguistics, Saarbrucken, Germany, 2000.
5. S. Varges. *"Instance-based Natural Language Generation"*. Ph.D. thesis, School of Informatics, University of Edinburgh, 2003.

6. A. Ratnaparkhi. *"Trainable approaches to surface natural language generation and their application to conversational dialogue systems"*. Computer, Speech & Language, 16(3):435–455, 2002.
7. A. Oh and A. Rudnicky. *"Stochastic natural language generation for spoken dialogue systems"*. Computer Speech & Language, pp. 387–407, 2002.
8. J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu and A. Stent. *"An architecture for a generic dialogue shell"*. Natural Language Engineering, 6(3), 2000.
9. A. Belz. *"Statistical generation: Three methods compared and evaluated"*. In Proceedings of the 10th European Workshop on Natural Language Generation, ENLG 05, 2005.
10. A. Mustapha, M. N. Sulaiman, R. Mahmod and M. H. Selamat. *"Classification-and-ranking architecture for response generation based on intentions"*. International Journal of Computer Science and Network Security, 8(12):253–258, 2008.
11. M. White. *"Reining in CCG chart realization"*. In Proceedings of the 3rd International Natural Language Generation Conference, Hampshire, UK, 2004.
12. E. Velldal and S. Oepen. *"Maximum entropy model for realization ranking"*. In Proceedings of the 10th MT-Summit (X), Phuket, Thailand, 2005.
13. G.v.d. Hoeven, A. Andernach, S.v.d. Burgt, G-J. Kruijff, A. Nijholt, J. Schaake and F.d. Jong. *"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, pp. 271-285, Versailles, France, 1995.
14. J. Hulstijn. *"Dialogue models for inquiry and transaction"*. Ph.D. thesis, University of Twente, Netherlands, 2000.
15. A. Stent. *"A conversation acts model for generating spoken dialogue contributions"*. Computer Speech and Language, 16(3):313–352, 2002.