Journal of Computer Science 6 (11): 1326-1333, 2010
ISSN 1549-3636
© 2010 Science Publications

Corpus-Based Analysis on Cross-Domain Experiments in Classification-and-Ranking Generation

Aida Mustapha, Md. Nasir Sulaiman, Ramlan Mahmod and Mohd. Hasan Selamat
Department of Computer Science, Faculty of Computer Science and Information Technology,
University Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia

Abstract: Problem statement: Overgeneration-and-ranking architecture works 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. **Approach:** In classification-and-ranking, response was 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. Because the architecture is intention-based, it adopted an open-domain knowledge representation, whereby response utterances were semantically represented using some ontology general enough for future reuse in another domain. **Results:** This study presented corpus-based analysis on cross-domain experimentation using different type of corpus to validate the consistency of the response classifier that delimits the searching space for ranking. The open-domain quality for classification-an-ranking architecture was tested on two mixed-initiative, transaction dialogue corpus in theater reservation and emergency planning. Results showed consistent distribution accuracies in both classification and ranking experiment, indicating that the approach is viable for cross-domain implementations. **Conclusion:** The ability of a response generation system to directly learn response utterances from the domain corpus suggested the possibility to build a dialogue system by feeding the learning module with a target corpus and the system learned the response behavior directly from the training corpus.

Key words: Corpus-based, open-domain, classification-and-ranking, natural language generation, dialogue systems

INTRODUCTION

Response generation is essentially the natural language generation component in dialogue systems. Many response generation systems are lacking robustness in implementation. This is attributed to the high degree requirement of linguistic specifications, which is manual construction of grammar rules to generate response utterances (Varges and Purver 2006; Ward, 1994). This problem has, in turn, motivated the statistical approach to automatically learn language models from the corpus so the response generation systems do not have to depend on grammar rules anymore. In overgeneration-and-ranking (Langkilde and Knight, 1998), utterances are generated either through minimal rule-based transformation or statistical language distribution in all possible combinations, including fragments. A separate ranking module is required to rank the candidate utterances and select the best response from the ranking stage.

Overgeneration-and-ranking architecture evades linguistic decision-making process of grammar and template-based (Varges and Purver, 2006; Langkilde and Knight, 1998; Bangalore and Rambow, 2000; Marciniak and Strube, 2004; Reiter and Mellish, 1992). Nonetheless, although generation is robust, it is expensive because learning the language model is performed statistically. This means, every alternative realization and its probabilities have to be calculated individually. Language models are also known to have built-in bias in producing short strings because the likelihood of a string is determined by joint probability of the words (Belz, 2005). As opposed to generation of sentences, this bias is not desirable for generation of dialogue utterances because utterances render intention in the form of speech actions. Therefore, all statistically realized utterances should be weighted based on intentions and should be treated as equally good realizations regardless of length, in fact, regardless of grammar.

Corresponding Author: Aida Mustapha, Department of Computer Science, Faculty of Computer Science and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia
Tel: +60-3-8947-1714 Fax: +60-3-8946-6577

To cater for short strings, classification-and-ranking architecture (Mustapha *et al*., 2008) advocates for intention-based response generation with independent domain representation. The philosophy behind the architecture is that the generation system learns to manage its own response strategies based on corpus. Possible responses are deliberately chosen from dialogue corpus rather than wholly generated, so the approach allows short ungrammatical utterances (fragments) as long as they satisfy the intended meaning of the input utterance. Because the architecture assumes that each user utterance represented in some context has its counterpart response, learning is robust and the response generation system should be able to cross domain.

Classification-and-ranking generation: Figure 1 shows the two-staged process in classification-and-ranking architecture. The first component is a Bayesian classifier to classify user utterances into response classes based on intentions of user input utterance. The second component is an entropic ranker that scores the candidate response utterances in each response class according to semantics relevant to the input utterance. The basic mechanic is to find the most coherent response class where the possible responses reside in. Next is to access the relevance of response candidates in that particular response class, which satisfies the intentions of user input utterance.

Classification of response utterances is necessary in order to delimit the searching space for ranking the utterances. Classification-and-ranking advocates for domain-independent generation system through knowledge abstraction of utterances using some domain ontology that is general enough for similar representation in another domain. Hence, the corpus-based learning is consequently challenged by the ability to model the domain contextual knowledge such that the knowledge encoded in each utterance may be ranked as more informative from one another. Previous works on open-domain dialogue systems focus on dialogue modeling through domain-independent semantic input for over generation and an out-of-domain language model like newswire or Penn Treebank for ranking (Chen *et al*., 2002; Chambers and Allen, 2004; Sharif and Saad, 2005).

Dialogue modeling: The rudimentary approach to modeling dialogue utterances lies in the way utterances are classified into in the first place, which is through topical contributions of the utterances. This means, utterances that are grouped together have topic as the common factor, but the relation of the topic to context varies. To distinguish between the utterances within the same topical classification, classification-and-ranking advocates for focus of attention (McKeown, 1985) to constrain the information that needs to be considered when deciding which response to realize. Because the dialogue utterances are classified into topics, therefore, focus will bind the response best to a particular input utterance.

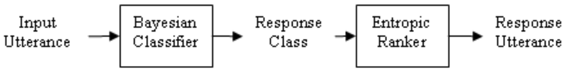


Fig. 1: Classification-and-ranking architecture

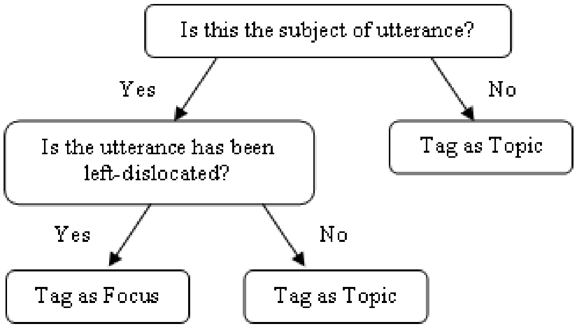

Is this the subject of utterance?
Yes
No
Is the utterance has been left-dislocated?
Tag as Topic
Yes
No
Tag as Focus
Tag as Topic

Fig. 2: Decision tree for topic and focus extraction

Nonetheless, the set of dialogue utterances classified under the same topic may communicate different meaning altogether, even when bounded with the focus. The second level of discrimination warrant for the worth of the knowledge itself encoded in each response utterances. This means utterances must be abstracted out into domain attributes so measuring in formativeness is possible. Extraction of topic and focus is based on Information Structure Theory (Halliday, 1967) that adopts topic articulation as the first element in the sentence. Hence, topics are accessed depending on the structure of utterances such as assertive, imperative and interrogative.

Figure 2 illustrates the decision tree to extract topic and focus. While information structure does not primarily affect the truth conditions of utterances, it does affect the packaging hence the emphasis of the utterances. Formalism of the domain attributes enable the domain to be extended or replaced, hence making the response generation system more robust and domain-independent.

MATERIALS AND METHODS

The objective of a secondary, cross-domain experimentation using a different body of corpus is to evaluate ability of the intention-based, classification-and-ranking response generation to classify and rank in another domain. Because the architecture assume the existence of dialogue act-annotated dialogue corpus based on DAMSL (Core and Allen, 1997) annotation scheme, the secondary corpus must have at least grounding and speech acts in order to run comparative experiment. The experiments described in this study are performed comparatively using two corpus, namely SCHISMA (Hoeven *et al*., 1995) and MONROE (Stent, 2000).

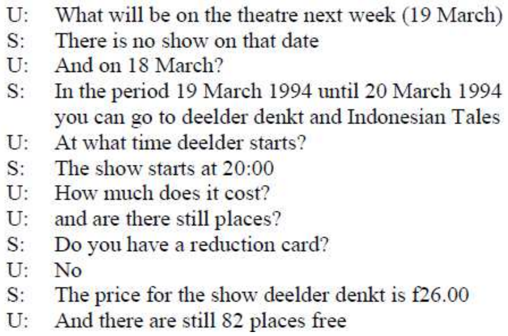
In sourcing the corpus, three existing corpora are dialogue-act annotated, which are SWITCHBOARD (Daniel *et al*., 1997), TRAINS-93 (Allen *et al*., 1995) and MONROE (Stent, 2000). While SWITCHBOARD corpus has been annotated with speech acts using a variant of DAMSL coding, only parts of TRAINS-93 corpus has been annotated with speech acts using the same annotation scheme. Nevertheless, MONROE corpus has been tagged with complete conversation act behaviors, which are turn-taking, grounding and speech acts, argumentation acts, as well as initiative.

SCHISMA corpus: The SCHISMA corpus provides insights into user behaviors and typical languages used in a limited theater reservation domain. Among the objects to perform theater reservation are title, date, genre and artist. The entire SCHISMA corpus is constituted by 64 dialogues of varied length, each consist of a single or a sequence of inquiry and transaction dialogues. SCHISMA is also readily annotated with conversational acts. Excerpts of dialogues in MONROE are illustrated in Fig. 3.

MONROE corpus: The MONROE corpus is a collaborative problem-solving task in disaster scenario set in Monroe County, New York. Disaster scenarios include car accidents, natural disasters like flood and snow storms, request for medical assistance, or civil disorders. Given a particular emergency task, the dialogue participants are expected to coordinate help for the task. Among the objects to coordinate for in this domain include people, roads, vehicles, crews and equipment. Because of the complexities in the collaboration task, the corpus is essentially longer with 1,568 lines of utterances even though it exists in eight dialogues. Excerpts of dialogues in MONROE are illustrated in Fig. 4.

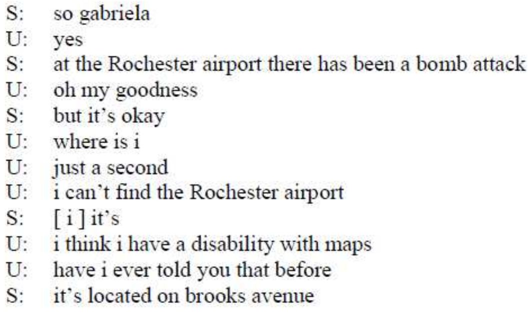
MONROE is similar to SCHISMA in terms of richness of the domain, whereby there are different types of objects that can be manipulated and actions that can be performed. However, MONROE is different from SCHISMA in three aspects. One, the dialogues are human-human, while SCHISMA is collected through a simulated Wizard-of-Oz experiment. Two, MONROE is speech-based, which means that the corpus is collected in audio files and annotation is through listening to the audio corpus. SCHISMA on the other hand is text-based. Three, the number of dialogues in MONROE is small in size (which is only eight dialogues) but are lengthy as compared to the small but compact 64 dialogues of SCHISMA. Table 1 shows statistical comparison between SCHISMA and MONROE corpus.

Bayesian classification experiment: The objective of response classification task in the classification-and-ranking architecture is to determine which of a set of response classes that the user input utterance belongs to. Response classification is performed using Bayesian networks to find the recognition accuracy of correct predictions for response class rc, given the user utterance U, or simply P(response class|user utterance).



U: What will be on the theatre next week (19 March)
S: There is no show on that date
U: And on 18 March?
S: In the period 19 March 1994 until 20 March 1994 you can go to deelder denkt and Indonesian Tales
U: At what time deelder starts?
S: The show starts at 20:00
U: How much does it cost?
U: and are there still places?
S: Do you have a reduction card?
U: No
S: The price for the show deelder denkt is f26.00
U: And there are still 82 places free

Fig. 3: Excerpt of dialogues in SCHISMA corpus



S: so gabriela
U: yes
S: at the Rochester airport there has been a bomb attack
U: oh my goodness
S: but it's okay
U: where is i
U: just a second
U: i can't find the Rochester airport
S: [i] it's
U: i think i have a disability with maps
U: have i ever told you that before
S: it's located on brooks avenue

Fig. 4: Excerpt of dialogues in MONROE corpus

Table 1: Statistical comparison between both corpora

Corpus	SCHISMA	MONROE
Number of dialogue	64.000	8.000
Number of utterance	2.047	2.899
Number of turns	1.723	1.758
Number of word	20.565	19.328

Table 2: Features represented as nodes in Bayesian networks

Features	Node	Type
Context	CX	Scalar
Topic	T	Scalar
Speech act	FLF	Scalar
Grounding act	BLF	Scalar
Mood	M	Scalar
Control	C	Scalar
Role	R	Scalar
Turn-taking act	TU	Scalar
Argumentation act	N	Scalar
Response class	RC	Scalar

Table 3: Response classes for SCHISMA and MONROE

SCHISMA	MONROE
Title	Emergency
Genre	Condition
Artist	Victim
Time	Time
Date	Distance
Review	Landmark
Person	Scene
Reserve	Map
Ticket	Vehicle
Cost	Crew
Avail	Location
Reduce	Equipment
Seat	Station
Theater	Hospital
Other	Other

The decision rule for the classification experiment is given Eq. 1, where $\hat{rc}$ is the estimate of the correct response class:

$$\hat{rc} = \underset{rc \in R}{\arg\max}\, P(U|rc)P(rc) \qquad (1)$$

Features from user utterances in both SCHISMA and MONROE corpus are characterized into two levels; semantic features and pragmatic features, which are represented as nodes in the Bayesian networks as shown in Table 2. Semantic features are (1) context that represents the global topic and (2) topic that represents the topic of user utterance. Pragmatic features are (1) speech act (FLF) that represents the intention of user; (2) grounding act (BLF) that represents the acknowledgement of user; (3) mood that states the linguistic mood of user utterance; (4) control that represents the party who holds the control initiative; (5) role that tells what the role of the system is at that particular utterance; (6) turn which represents the turn-taking act in user utterance and (9) argumentation that represents the argumentation acts in the current utterance.

Given the set of semantic and pragmatic features in each user input utterance, the response classes rc are being manually tagged according to the topic of the utterances. Tagging faithfully adapts to patterns of adjacency pairs from input and response utterance per turn throughout the course of conversation. While division of response classes maintain at the same quantity, however, the response classes are qualitative. Therefore, they are unique from corpus to corpus. Table 3 shows the response classes for both SCHISMA and MONROE corpus.

The classification experiment is divided into two cases. Case 1 is using conversation acts features while Case 2 is exploring time-series features. The goal Case 1 is to investigate the impact of intentions in an utterance under the interpretation of a conversational framework, based on Conversation Acts Theory (Traum and Hinkelman, 1992). This theory enables dialogue modeling to capture interaction of intentions at all levels during communication. The theory distinguishes four levels of action that are necessary for maintaining coherence and content of conversation, which are speech acts, grounding acts, turn-taking acts and argumentation acts.

The goal of Case 2 is to investigate the impact of features extracted from previous n user input utterances, if the semantic or pragmatic representation from earlier conversation has any influence over the accuracy rate in classification of response utterance. Classification on both SCHISMA and MONROE is using the same set of features but with values specifically extracted from the respective corpus. For each case, a 10-fold cross-validation is performed to split the data into training and testing sets.

Entropic ranking: Having identified the correct response class, second stage of the intention-based, classification and-ranking architecture is to identify the best response utterance within the particular class. Ranking is performed separately for each response class from the Bayesian classification module, specific to SCHISMA and MONROE.

The response class rc holds a set of possible response utterances $\{r_1 r_2 \ldots r_R\}$ from the repertoire of responses R. The goal of the ranking module is to output a single response utterance $r \in \{r_1 r_2 \ldots r_R\}$ in respond to the user; by choosing the response with the highest probability score.

The probability model is defined over R × S, where R is the set of possible response utterances and S is the set of corresponding semantic features to each response utterance. 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, while global knowledge is supplied by semantic input from user utterance. The features are described in Table 4.

Table 4: Local and global knowledge for R

S	Features	Descriptions
Local	rtopic	Topic of conversation in response utterance
Local	rfocus	Focus of attention in response utterance
Local	rflf	Speech act for response utterance
Local	rblf	Grounding act for response utterance
Local	da	Domain attributes in response utterance
Global	ufocus	Focus of attention in user utterance

Using both local and global features to model the probability distribution, each evidence in the training data has M feature functions $f_m(r, \{r_1r_2\ldots r_R\}, s)$ where $r\in R$, $s\in S$ and $m = 1, \ldots, M$. The probability model for our Entropic ranking of response utterance r conditioned to features s is defined as Eq. 2, where λ_m is the weights associated with each feature m and the normalizing function Z(s) is defined in Eq. 3. Given this modeling equation, we arrive at the decision rule in Eq. 4:

$$p(r\left|\{r_1\,r_2\ldots r_R\},s\right) = \frac{1}{Z(s)}\exp\left[\sum_{m=1}^{M}\lambda_m f_m(r,(\{r_1\,r_2\ldots r_R\},s\right] \qquad (2)$$

$$Z(s) = \sum\nolimits_{r'}\exp\left[\sum_{m=1}^{M}\lambda_m f_m(r',(\{r_1\,r_2\ldots r_R\},s\right] \qquad (3)$$

$$\begin{aligned}\hat{r} &= \arg\max_{r\in R}\left[p(r\mid\{r_1r_2\ldots r_R\},s)\right]\\ &= \arg\max_{r\in R}\left[\sum_{m=1}^{M}\lambda_m f_m(r,\{r_1r_2\ldots r_R\},s)\right]\end{aligned} \qquad (4)$$

The maximum entropy model will rank all the response utterances according to local and global features from response utterances as listed in Table 4. Assuming the response class rc supplied by the classification module is correct, training and testing was performed separately for all 15 response classes as shown in Table 3.

RESULTS

Bayesian classification: The baseline accuracy for the classification experiment is the majority baseline, which is taken from relative frequency of the most frequent response class. The baseline for SCHISMA is 16.3% coming from the class reserve. The baseline for MONROE is 22.3% coming from the class vehicle. Table 5 shows the results for response classification experiments using different set of features for both SCHISMA and MONROE.

Note that experiments in Case 2 are considering semantic and pragmatic features from current utterance up to three levels of previous utterances, which are represented as topic-n or FLF-n for semantic content and intentions, respectively.

Entropic ranking: The baseline accuracy to randomly pick a response utterance is 21.8% in SCHISMA and 23.8% in MONROE. Table 6 shows the results for ranking experiment in response classes. Note that response classes are topical and specific to each corpus.

DISCUSSION

Table 7 summarizes the result for both classification and ranking experiments using two sets of corpus, SCHISMA and MONROE. The results for classification cases in MONROE, although lower, are consistent to results in SCHISMA. The first factor that contributes to the differences in accuracy percentage is the size of the dialogue corpus. In the case of MONROE, albeit the 6% deviation in word counts of SCHISMA and MONROE, SCHISMA is made up of 64 dialogues compared to 8 dialogues of MONROE. The high number of dialogues indicates that the dialogues are shorter, more compact, hence the extraction of semantics and pragmatics features are more accurate. In MONROE, dialogues are stretched to a longer span; hence include distortions in terms of out of topic discussion or a lengthy explanation that carry the same semantic information, thus insignificant in context. Consistent with results from SCHISMA, results from the experiment show that time-series experiments from Case 2 in MONROE do not significantly improve the classification accuracy as shown in Table 5. In addition, regardless the increase, the accuracy rate will deteriorate as we add more utterances from the history of conversation. This means incorporating more features from previous utterances only increase the recognition accuracy insignificantly. The most important observation is the accuracy rates tend to fall once the features are dragged across longer span of dialogue utterances.

As for the ranking experiment, while division of response classes maintain at the same quantity, however, the classes are unique for both SCHISMA and MONROE. The lower ranking accuracies from MONROE corpus as shown in Table 6 are consistent with the distribution of results in SCHISMA corpus. This can be seen from Fig. 5, which illustrates the accuracy distribution for response classes in both corpora. The consistent distribution pattern shows that intention-based classification-and-ranking response generation is able to cross domain.

Classification-and-ranking generation provides a principled way to combine pragmatic interpretations of user utterance and informativeness of the response utterance as the basis for knowledge abstraction. It also takes over the responsibility of dialogue manager through intention-based dialogue modeling. This approach sidesteps the entire generation but instead performs corpus-based learning through classification.

Table 5: Comparative results for response classification

		Accuracy (%)	
Semantic features	Pragmatic features	SCHISMA	MONROE
Context, Topic	Case 2, Negotiation	73.9	64.8
Context, Topic	Turn, FLF, BLF, Negotiation	74.0	64.8
Context, Topic, Topic-1	Turn, FLF, BLF, Negotiation	74.1	64.9
Context, Topic, Topic-2	Turn, FLF, BLF, Negotiation	73.2	64.9
Context, Topic, Topic-3	Turn, FLF, BLF, Negotiation	72.8	64.7
Context, Topic	Turn, FLF, BLF, Negotiation, FLF-1	73.6	65.0
Context, Topic	Turn, FLF, BLF, Negotiation, FLF-2	73.8	64.6
Context, Topic	Turn, FLF, BLF, Negotiation, FLF-3	73.4	62.4

Table 6: Comparative results for ranking

	SCHISMA			MONROE			
Response class	No. of instances	Error rate	Accuracy (%)	Response class	No. of instances	Error rate	Accuracy (%)
Title	104	9	91.3000	Emergency	148	24	83.8
Genre	28	1	89.3000	Condition	21	3	85.7
Artist	42	4	90.5000	Victim	141	19	86.5
Time	32	3	90.6000	Time	78	16	79.5
Date	90	13	93.3000	Distance	19	0	100.0
Review	56	5	89.3000	Landmark	6	1	83.3
Person	30	3	96.7000	Scene	270	77	71.5
Reserve	15	0	31 79.30	Map	20	0	100.0
Ticket	81	2	97.5000	Vehicle	350	122	65.1
Cost	53	4	90.6000	Crew	270	75	72.2
Avail	14	1	92.9000	Location	19	2	89.5
Reduce	73	3	93.2000	Equipment	53	5	90.6
Seat	94	6	93.6000	Station	8	1	87.5
Theater	12	1	100.0000	Hospital	53	9	83.0
Other	61	10	80.3000	Other	112	22	80.4
	920		91.2		1,568		84.0

Table 7: Summary of results

Task	SCHISMA (%)	MONROE (%)
Classification	74.0	64.8
Ranking	91.2	84.0

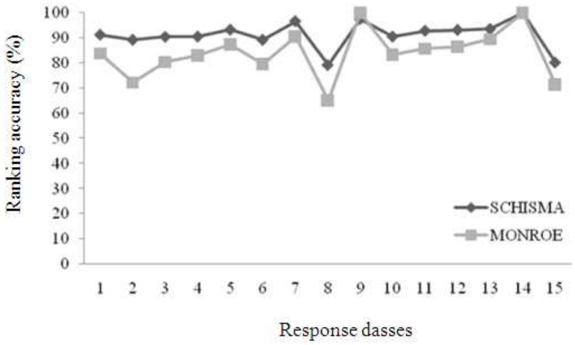


Fig. 5: Distribution of ranking accuracies between SCHISMA and MONROE

Although a ranking component exists, the ranker ranks the response utterance on basis of relevance with regards to the input, therefore carries more weight as the resulting utterance must be coherent.

The main problem with recognition accuracy lies in features extractions, whose successful rate highly depends on size of the dialogue corpus. In the case of MONROE, albeit the 6% deviation in word counts between both corpus, SCHISMA is made up of 64 dialogues as compared to 8 dialogues in MONROE. The high number of dialogues indicates that the dialogues are shorter, more compact, hence the extraction of semantics and pragmatics features are more accurate. In MONROE, however, dialogues are stretched to a longer span, hence may include distortions in terms of out of topic discussion or a lengthy explanation that carry the same semantic information, thus insignificant in terms of context.

MONROE corpus is also a speech-based, human-human conversation. A number of complications may arise in tagging the initial dialogue acts alone, for instance, frequent overlaps lead to annotators' disagreement on tagging the dialogue acts are reported in (Stent, 2000). Distribution of semantic features is often skewed because human-human conversation tends to have many fillers and irrelevant remarks, including continuers or backchannels like "hmm" and "uh-uh". A mutual human-human conversation also tends to exhibit more adjacency pairs of assertion-agreement rather than

Question-Answer as in human-machine conversation. However, the size of dialogue corpus is of bigger impact as compared to the nature of the corpus. This is because the intention-based approach relies on semantic and pragmatic features extraction that is independent from the form of surface utterance, whether the utterances are human-machine or human-human.

CONCLUSION

The central focus of this study is to present analysis on cross-domain experiment for classification-and-ranking response generation. The approach to domain-independent generation system is through knowledge abstraction using domain ontology general enough for future reuse in another domain, for example title and genre in SCHISMA against vehicle and crew in MONROE. Again, the response classes used in classification experiment are qualitative; hence are independent from one corpus to another.

Given a set of utterances in a corpus, this study concludes that a classification-and-ranking generation system is able to discriminate among utterances in a specific response class based on equality of informativeness in the utterances. The notion of “equivalent” rather than “identical” is based on the ground that there could be more than one utterance that conveys the same semantic and pragmatic interpretation but exists in different form of linguistic structures. An equivalent response also reflects that the response is coherent to the dialogue context and relevant to the preceding input utterance.

The ability of a response generation system to directly learn response utterances from the domain corpus suggests the possibility to build a dialogue system by feeding the learning module with a target corpus and the system learns the response behavior directly from the training corpus. This will enable the system to jumpstart a conversation in a particular domain and allow system engineers to transform the system behavior only by changing corpora. In turn, this will avoid the time-consuming and labor-intensive manual response construction when a response generation system is ported to a new domain. Among domains that could benefit the classification-and ranking architecture include information-based systems like customer service for credit cards, ticket reservation and transportation scheduling.

In the future, this research will branch into two directions. The first direction is to investigate the cross-domain experiment in Malays corpus in (Tan and Salleh, 2009). The second direction is to investigate the performance of Bayesian networks during ranking as opposed to classification based on Saat *et al*. (2010).

REFERENCES

Allen, J.F., L.K. Schubert, G. Ferguson, P. Heeman and C.H. Hwang *et al*., 1995. The TRAINS project: A case study in building a conversational planning agent. J. Exp. Theor. Artif. Intell., 7: 7-48. DOI: 10.1080/09528139508953799

Bangalore, S. and O. Rambow, 2000. Exploiting a probabilistic hierarchical model for generation. Proceeding of the 18th International Conference on Computational Linguistics, July 31-Aug. 4, Association for Computational Linguistics, Saarbrucken, Germany, pp: 42-48. DOI: 10.3115/990820.990827

Belz, A., 2005. Statistical generation: Three methods compared and evaluated. Proceeding of the 10th European Workshop on Natural Language Generation, Aug. 8-10, CiteSeerX, USA., pp: 15-23. http://www.itri.brighton.ac.uk/~Anja.Belz/Publications/belz-enlg05.pdf

Chambers, N. and J. Allen, 2004. Stochastic language generation in a dialogue system: Toward a domain independent generator. Proceeding of the 5th SIGdial Workshop on Discourse and Dialogue, Apr. 30-May 1, Association for Computational Linguistics, Boston, USA., pp: 1-11.

Chen, J., S. Bangalore, O. Rambow and M.A. Walker, 2002. Towards automatic generation of natural language generation systems. Proceeding of the 19th International Conference on Computational Linguistics, Aug. 24-Sept. 1, Association for Computational Linguistics, Taipei, Taiwan, pp: 1-7. DOI: 10.3115/1072228.1072366

Core, M.G. and J.F. Allen, 1997. Coding dialogs with the DAMSL annotation scheme. Proceeding of the Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines, Nov. 8-10, CiteSeerX, Boston, MA., pp: 28-35.

Daniel, J., R. Bates, N. Coccaro, R. Martin and M. Meteer *et al*., 1997. Switchboard discourse language modeling project report. Center for Speech and Language Processing. http://www.bibsonomy.org/bibtex/26dbd915d0585f37c37c40bb97f4dc797/sonntag?layout=plain

Halliday, M.A.K., 1967. Notes on transitivity and theme in English: Part 2. J. Linguist., 3: 199-244. http://www.jstor.org/pss/4174965

Hoeven, G.F.V.D., G.A. Andernach, S.V. van de Burgt, G.J.M. Kruijff and A. Nijholt *et al*., 1995. SCHISMA: A natural language accessible theatre information and booking system. Proceeding of the 1st International Workshop on Applications of Natural Language to Data Bases, June 28-29, PubZone, Versailles, France, pp: 271-285.

Langkilde, I. and K. Knight, 1998. Generation that exploits corpus-based statistical knowledge. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, Aug. 10-14, Association for Computational Linguistics, Montreal, Quebec, Canada, pp: 704-710. DOI: 10.3115/980845.980963

Marciniak, T. and M. Strube, 2004. Classification-based generation using TAG. Lecture Notes Comput. Sci., 3123: 100-109. DOI: 10.1007/978-3-540-27823-8

McKeown, K., 1985. Text generation (Studies in Natural Language Processing). 1st Edn., Cambridge University Press, Cambridge, ISBN: 10: 0521301165, pp: 264.

Mustapha, A. M.N. Sulaiman, R. Mahmod and H. Selamat, 2008. Classification-and-ranking architecture for response generation based on intentions. Int. J. Comput. Sci. Network Secur., 8: 253-258. http://paper.ijcsns.org/07_book/200812/20081236.pdf

Reiter, E. and C. Mellish, 1992. Using classification to generate text. Proceeding of the 30th Annual Meeting on Association for Computational Linguistics, June 28-July 2, Association for Computational Linguistics, Newark, Delaware, pp: 265-272. DOI: 10.3115/981967.982001

Saat, N.Z.M., K. Ibrahim and A.A. Jemain, 2010. Bayesian methods for ranking the severity of apnea among patients. Am. J. Applied Sci., 7: 167-170. DOI: 10.3844/ajassp.2010.167.170

Stent, A.J., 2000. The Monroe corpus. University of Rochester. http://portal.acm.org/citation.cfm?id=898497

Traum, D.R. and E.A. Hinkelman, 1992. Conversation acts in task-oriented spoken dialogue. Comput. Intel., 8: 575-99. DOI: 10.1111/j.1467-8640.1992.tb00380.x

Tan, T.S. and S. Hussain, 2009. Corpus design for Malay corpus-based speech synthesis system. Am. J. Applied Sci., 6: 696-702. DOI: 10.3844/ajassp.2009.696.702

Varges, S. and M. Purver, 2006. Robust language analysis and generation for spoken dialogue systems. Proceeding of the 17th European Conference on Artificial Intelligence, Workshop on Development and Evaluation of Robust Spoken Dialogue Systems for Real Applications, Aug. 29-Sept. 1, Riva del Garda, Trentino, Italy, pp: 1-4.

Ward, N., 1994. A Connectionist Language Generator. Ablex Pub., Norwood, New Jersey, ISBN: 10: 1567500382, pp: 310.