



UNIVERSITI PUTRA MALAYSIA

**CLASSIFICATION-AND-RANKING ARCHITECTURE BASED ON
INTENTIONS FOR RESPONSE GENERATION SYSTEMS**

AIDA MUSTAPHA.

FSKTM 2008 1



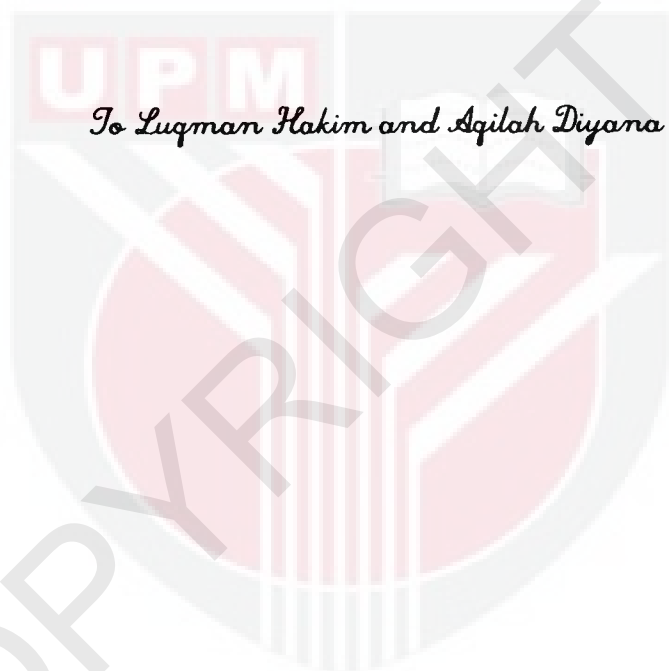
**CLASSIFICATION-AND-RANKING ARCHITECTURE BASED ON INTENTIONS
FOR RESPONSE GENERATION SYSTEMS**

By
AIDA MUSTAPHA

**Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia,
in Fulfilment of the Requirement for the Degree of Doctor of Philosophy**

March 2008





To Luqman Hakim and Aqilah Diyana



Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

**CLASSIFICATION-AND-RANKING ARCHITECTURE BASED ON INTENTIONS
FOR RESPONSE GENERATION SYSTEMS**

By

AIDA MUSTAPHA

March 2008

Chairman: Associate Professor Md. Nasir Sulaiman, PhD

Faculty: Computer Science and Information Technology

Existing response generation accounts only concern with generation of words into sentences, either by means of grammar or statistical distribution. While the resulting utterance may be inarguably sophisticated, the impact may be not 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 the response robustly satisfies the intention of input utterance. At the same time the response must maintain coherence and relevance, regardless of the surface presentation. This means that response generation is constrained by the content of intentions, rather than the lexicons and grammar.

Statistical techniques, mainly the 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. The bias towards short

strings during ranking is the very limitation of this approach hence leading to our proposed intention-based classification-and-ranking architecture.

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 architecture employs two basic components, which is a Bayesian classifier to classify user utterances into response classes based on their pragmatic interpretations, and an Entropic ranker that scores the candidate response utterances according to the semantic content relevant to the user utterance. The high-level, pragmatic knowledge in user utterances are used as features in Bayesian classification to constrain response utterance according to their contextual contributions, therefore, guiding our Maximum Entropy ranking process to find one single response utterance that is most relevant to the input utterance.

The proposed architecture is tested on a mixed-initiative, transaction dialogue corpus of 64 conversations in theater information and reservation system. We measure the output of the intention-based response generation based on coherence of the response against the input utterance in the test set. We also tested the architecture on the second body of corpus in emergency planning to warrant the portability of architecture to cross domains. In the essence, intention-based response generation performs better as compared to surface generation because features used in the architecture extend well into pragmatics, beyond the linguistic forms and semantic interpretations.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**SENI BINA PENGELASAN-DAN-PENGATURAN BERASASKAN MAKSUD
TERSIRAT BAGI SISTEM PENJANAAN RESPON**

Oleh

AIDA MUSTAPHA

Mac 2008

Pengerusi: Profesor Madya Md. Nasir Sulaiman, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

Huraian tentang penjanaaan respon yang sedia ada hanya menitikberatkan penjanaaan perkataan ke dalam bentuk ayat, sama ada melalui petua-petua nahu atau taburan statistik. Walaupun ayat-ayat yang dijanakan terbukti sofistikated, ianya kurang memberi kesan. Kami percaya bahawa reka bentuk bagi penjanaaan respon memerlukan lebih daripada petua nahu atau taburan statistik, tetapi lebih berintuisi atau jelas, relevan dan menjawab kepada kehendak pertuturan yang tersirat. Ini bermakna, penjanaaan respon adalah tertakluk kepada kekangan maksud tersirat dan bukannya abjad atau nahu bahasa.

Teknik-teknik statistik terutamanya seni bina penjanaaan-dan-pengaturan berfungsi dengan baik ke atas bahasa penulisan, di mana ayat merupakan unit asas. Dalam bahasa pertuturan, walau bagaimanapun, ayat yang dituturkan tidak sama dengan ayat yang ditulis kerana maksud tersirat turut memainkan peranan di samping membenarkan penggunaan ayat-ayat pendek asalkan pembawaan maksud yang sama. Pengaruh kecenderungan ini merupakan

kekangan sebenar kepada seni bina tersebut dan membawa kepada cadangan seni bina pengelasan-dan-pengaturan kami yang berasaskan maksud tersirat.

Dalam seni bina yang dicadangkan, respon tidak dijana secara langsung tetapi dipilih daripada sesebuah kumpulan dialog bagi membenarkan pemilihan respon pertuturan yang pendek sekalipun mengandungi kesalahan tatabahasa, asalkan respon tersebut menjawab kepada maksud tersirat. Seni bina ini mempunyai dua komponen asas, iaitu pengkelas Bayesian yang bertujuan untuk mengelaskan pertuturan input kepada kelas-kelas respon berdasarkan pemahaman pragmatik; dan pengatur Entropi yang membezakan ayat-ayat respon dalam setiap kelas tersebut berdasarkan kepada kandungan semantik respon-respon. Pengetahuan pragmatik dalam pertuturan input merupakan ciri utama dalam pengkelas Bayesian bagi mengekang pertuturan respon berdasarkan pengagihan konteks. Pada akhirnya, proses pengatur Entropi berupaya memilih hanya satu pertuturan respon yang dianggap relevan dan menjawab pertuturan input.

Seni bina yang dicadangkan telah diuji ke atas sebuah kumpulan dialog transaksi berbilang inisiatif yang terdiri daripada 64 perbualan dalam sebuah sistem maklumat dan tempahan teater. Kami membandingkan hasil penjaanan respon berdasarkan maksud tersirat ini dengan pertuturan input dalam set pengujian. Kami juga turut menguji seni bina yang dicadangkan ke atas kumpulan dialog kedua dalam bidang perancangan kecemasan bagi menjamin ciri mudah alih seni bina tersebut ke dalam bidang pengetahuan yang lain. Secara asasnya, penjaanan respon berdasarkan maksud tersirat adalah lebih baik daripada penjaanan respon berdasarkan ciri-ciri linguistik kerana ciri yang digunapakai menjangkau aras pragmatik dan di luar batasan linguistik dan pemahaman semantik.

ACKNOWLEDGEMENTS

Profound thanks to Assoc. Prof. Dr. Md. Nasir Sulaiman for shaping the research in his myriad of questions, deepest sense of gratitude to Assoc. Prof. Dr. Ramlan Mahmud for his meticulous eyes on details, and Assoc. Prof. Mohd. Hasan Selamat for being available at times when things did not seem to work as planned. Thanks are also due to Anwar Ali Yahya for valuable long-labored discussions, and all friends at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia for providing a welcome distraction from the research.

TABLE OF CONTENTS

	Page
DEDICATION	ii
ABSTRACT	iii
ABSTRAK	v
ACKNOWLEDGEMENTS	vii
APPROVAL	viii
DECLARATION	x
LIST OF TABLES	xiv
LIST OF FIGURES	xvi
LIST OF ABBREVIATIONS	xviii
CHAPTER	
1 INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	4
1.3 Objectives of research	4
1.4 Scope of research	6
1.5 Research methodology	6
1.6 Organization of this thesis	9
2 LITERATURE REVIEWS	11
2.1 Introduction	11
2.2 Natural language generation (NLG)	13
2.2.1 Tasks and architecture	13
2.2.2 Deep generation	16
2.2.3 Surface generation	19
2.3 NLG specific to dialogues	25
2.3.1 Dialogue modeling	27
2.3.2 Grammar vs. templates	28
2.3.3 Corpus-based approach	30
2.3.4 Overgeneration-and-ranking	31
2.3.5 Information-state generation	32
2.4 Summary	35
3 THEORETICAL BACKGROUND	36
3.1 Introduction	36
3.2 Bayesian networks (BN)	39
3.2.2 Network structure	41
3.2.3 Conditional probability distributions	43
3.2.4 Inference	46
3.2.5 Learning	48



3.3	Dynamic Bayesian networks (DBN)	50
3.3.1	Network structure	51
3.3.2	Inference	53
3.3.3	Learning	55
3.4	Maximum Entropy (ME)	56
3.4.1	Entropy measures	59
3.4.2	Feature functions	60
3.4.3	Parameter learning	63
3.5	Summary	65
4	INTENTION-BASED RESPONSE GENERATION	67
4.1	Introduction	67
4.2	Speech and intention	68
4.3	Intention-based architecture	70
4.3.1	Intermediate representations	73
4.3.2	Response classification	74
4.3.3	Ranking utterances	74
4.3.4	Dialogue modeling	75
4.4	Foundation to Bayesian classification	77
4.4.1	Modeling mixed-initiative dialogues	77
4.4.2	Conversational act theory	78
4.4.3	Measuring coherence	85
4.5	Foundation to Entropic ranking	87
4.5.1	Modeling open-domain dialogues	88
4.5.2	Information structure theory	89
4.5.3	Measuring informativeness	93
4.6	Summary	96
5	FEATURES EXTRACTION	97
5.1	Introduction	97
5.2	About dialogues	98
5.3	SCHISMA corpus	101
5.4	DAMSL annotation scheme	104
5.4.1	Information-level	106
5.4.2	Forward-looking functions	107
5.4.3	Backward-looking functions	110
5.5	Extraction of semantic features	113
5.5.1	Topic and focus	115
5.5.2	Domain attributes	118
5.6	Extraction of pragmatic features	119
5.6.1	Mood of utterances	120
5.6.2	Control and initiatives	121
5.6.3	Intentions and grounding	123
5.6.4	Turn-taking	124
5.6.5	Argumentation	127
5.7	Tagging of response classes	130
5.8	Summary	132

6	RESULTS AND DISCUSSIONS	133
6.1	Introduction	133
6.2	Implementation of Bayesian classification	134
6.2.1	Experimental results	139
6.2.2	Extending to dynamic Bayesian networks	143
6.3	Implementation of Entropic ranking	146
6.3.1	Experimental results	149
6.3.2	Relationship to Maximum Likelihood	152
6.4	Comparison to knowledge-based approach	157
6.5	Comparison to overgeneration-and-ranking	159
6.5.1	Language models	160
6.5.2	Maximum entropy with language models	164
6.5.3	Instance-based learning	166
6.5.4	Discussions	169
6.6	Architectural comparison	172
6.7	Cross-domain validation	175
6.7.1	The MONROE corpus	176
6.7.2	Bayesian classification	178
6.7.3	Entropic ranking	180
6.7.4	Discussions	181
6.8	Summary	184
7	CONCLUSIONS AND FUTURE WORKS	185
7.1	Introduction	185
7.2	Research Contributions	187
7.3	Conclusion	187
7.3.1	Feature modeling	188
7.3.2	Bayesian classification	188
7.3.3	Entropic ranking	189
7.3.4	Cross-domain validation	189
7.4	Observations	190
7.5	Recommendations for future works	191

REFERENCES

APPENDICES

BIODATA OF STUDENT

LIST OF PUBLICATIONS

LIST OF TABLES

Table	Page
3.1 Nodes and values for SCHISMA-1	42
4.1 Conversation act types	79
4.2 Plan-based intentions for determining coherence	86
4.3 Syntactic location for subjects	91
5.1 Dialogue statistics for SCHISMA	104
5.2 Information-level (<i>IL</i>) for SCHISMA	106
5.3 Forward-looking functions (<i>FLF</i>) for SCHISMA	108
5.4 Backward-looking functions (<i>BLF</i>) for SCHISMA	111
5.5 Pairs of <i>FLF</i> and <i>BLF</i> for a single turn	112
5.6 Semantic features for SCHISMA	114
5.7 Mood of utterances in SCHISMA	120
5.8 <i>Control</i> and <i>roles</i> of utterances in SCHISMA	122
5.9 Turn-taking features for SCHISMA	125
5.10 Dialogue to demonstrate turn-taking	127
5.11 Negotiation features in argumentation level	128
5.12 Tagging of response classes to user utterance	130
5.13 Statistics for response classes	132
6.1 Nodes and values for response classification problem	135
6.2 Results for Case 1	140
6.3 Results for Case 2	141
6.4 Results for Case 3	143
6.5 Results for Case 4	144
6.6 Local and global knowledge for <i>R</i>	147
6.7 Features for Entropic ranking	148
6.8 Individual accuracy percentage for response classes in SCHISMA	150
6.9 Comparing ranking accuracy between ME and MLE	155
6.10 Statistical comparison between SCHISMA and MONROE	177
6.11 Results for Case 1	178



6.12	Results for Case 2	179
6.13	Individual accuracy percentage for response classes in MONROE	180
6.14	Classification and ranking results for SCHISMA and MONROE	181



LIST OF FIGURES

Figure	Page
1.1 Research methodology	8
2.1 Reference architecture for NLG system	15
2.2 Integration of NLG architecture into dialogue system	26
3.1 A BN for SCHISMA-1	42
3.2 A subset of BN for SCHISMA-1	45
3.3 Extending SCHISMA-1 into a temporal dimension	53
3.4 Types of inference	54
4.1 The two-staged classification-and-ranking architecture	72
4.2 Intention-based response generation	72
4.3 Input frame for user input utterance	73
4.4 Theoretical base to intention-based response generation	76
4.5 Argumentation acts in negotiation phases	84
4.6 Response utterances in response class <i>date</i>	94
5.1 A dialogue in SCHISMA	99
5.2 Decision tree for <i>FLF Statement</i>	109
5.3 Decision tree for <i>BLF Agreement</i>	111
5.4 Decision tree for <i>Topic and Focus</i>	115
5.5 Decision tree for <i>Mood</i> of utterances	121
5.6 Decision tree for <i>Control</i>	123
5.7 Decision tree for <i>Turn</i>	126
5.8 Decision tree for <i>Negotiation</i>	129
6.1 Bayesian networks for SCHISMA domain	138
6.2 Recognition accuracy by extending semantic content	145
6.3 Recognition accuracy by extending intentions	145
6.4 Ranking accuracy for SCHISMA	150
6.5 Response utterances in response class <i>reserve</i>	151
6.6 Response utterances in response class <i>other</i>	151
6.7 Response utterances in response class <i>date</i>	153



6.8	<u>Comparison between ME and MLE</u>	<u>156</u>
6.9	Knowledge-based NLG	157
6.10	Ranking using trigram LM	163
6.11	Response utterances in response class <i>genre</i>	164
6.12	Response utterances in response class <i>review</i>	164
6.13	Ranking using ME augmented with bigram LM	166
6.14	Ranking using instance-base learning (IBL)	168
6.15	Comparison of proposed approach against existing approaches	170
6.16	Comparison of surface-based approaches against MLE	171
6.17	Statistical-based (corpus-based) NLG	172
6.18	Intention-based NLG	174
6.19	Extending semantic contents in MONROE	179
6.20	Extending intentions in MONROE	180
6.21	Ranking accuracies for MONROE	181

LIST OF ABBREVIATIONS

BN	Bayesian Networks
CAT	Conversation Acts Theory
CPD	Conditional Probability Distribution
CPT	Conditional Probability Table
DBN	Dynamic Bayesian Networks
DM	Dialogue Manager
DST	Discourse Structure Theory
IBL	Instance-based Learning
JPD	Joint Probability Distribution
LM	Language Model
LPD	Local Probability Distribution
ME	Maximum Entropy
MLE	Maximum Likelihood Estimation
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
RST	Rhetorical Structure Theory



CHAPTER 1

INTRODUCTION

This chapter forms the introduction to the thesis. Discussions begin with emphasis of intentionality in speech and problems of generating responses for mixed-initiative, transaction-based dialogues under stochastic methodology.

1.1 Background

In human-human conversation, dialogue is mutually structured and timely negotiated between dialogue participants. Speakers take turns when they interact, they interrupt each other but their speeches seldom overlap. Each speaker is affected by what the other speaker has said, and what each speaker says, affect what the next speaker will say. Similarly, human-machine conversation through dialogue systems must exhibit comparable qualities. But for dialogue system to recognize turns, consider interrupts, and maintain coherence, response generation must rely on pragmatic interpretation, apart from semantic understanding of user input utterances.

Response generation is the Natural Language Generation (NLG) component in dialogue systems, which is responsible to construct the surface realization of the response utterances. In single-initiative dialogues system, there is an unequal balance of control. This often signifies some hierarchy of power whereby one party does all the asking and the other does all the answering (i.e., expert system, tutoring system). Generating responses for

such system is more structured and predictable because interruptions are not permissible. However, in collaboration type of dialogues (i.e., task-oriented system, planning system), both human and computer participants are working together to achieve common objectives. Because the interest is mutual, each participant shares equal balance of control, hence a mixed-initiative interaction (Hearst 1999).

A mixed-initiative dialogue system does not have a predetermined sequence of exchange structure. The party that initializes the dialogue has only temporary control over the initiative because the control is shared among participants. When change of initiatives is permitted, interrupts can happen, thus the flow of control transfer may be reversed by force. Additional complication arises when the task can only be achieved through negotiations, for instance in transaction dialogues. Disagreements, abandoned goals, and repetitive negotiations are all common before both participants finally accomplish the task and conclude the transaction (Hulstijn 2000).

Adapting mixed-initiative and transaction-based dialogues to statistical generation requires hybrid methodology drawn from the fields of both natural language generation (NLG) and pragmatics. NLG investigates how computer programs can be made to produce high-quality natural language text from underlying representations of meaning. The process is two-staged, deciding *what to say* (deep generation) and deciding *how to say* (surface realization) (Reiter and Dale 1997). Although surface generation has benefited from the robust corpus-based methodology, deep generation mostly remains the elegant grammar-based. However, neither approach has treated NLG through from the perspective of empirical pragmatics.

While semantics is the study of meaning in an utterance, pragmatics is the study of contribution of context to the utterance, thus providing a higher level account to interpretations. Semantics only encode the information within the utterance but pragmatic information can only be made relevant by the act of uttering the utterance (Bach 2002), which is through intentions.

As opposed to research on text generation systems that generate paragraph-length sentences, response generation in dialogue systems avoid detailed linguistic realization for 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 and still remains meaningful. Secondly, each dialogue utterance bears individual intention. Because utterances are intention-driven, the merit of an utterance depends on the magnitude of influence that the utterance imparts to the responding utterance. Given this, we hypothesize that even the surface form is grammatically incorrect, a response fares well as long as it satisfies the intentions of the utterance it is responding to.

However, the traditional approach to generating surface form of utterances is grammar-based hence is lacking robustness in implementation and is virtually incapable for any learning. The high degree requirement of linguistic specifications is the classic problem of knowledge engineering bottleneck (Ward 1994; Vargas 2003). This problem has, in turn, motivated for statistical approach to learn language models automatically so system does not have to depend on grammar rules anymore.

Nonetheless, statistical surface generation through language models, although robust, is expensive because alternative realizations and their probabilities have to be calculated individually. Furthermore, language models 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 (Belz 2005). This is clearly not desirable for generation of dialogue utterances because generation is real-time and all realized utterances should be treated as equally good realizations regardless of length, in fact, regardless of grammar.

1.2 Problem Statement

Existing response generation architecture only concern with generation of words into sentences, either by means of grammar or language model (Langkilde and Knight 1998; Bangalore and Rambow 2000; Langkilde 2000; Oh and Rudnicky 2000; Ratnaparkhi 2002; Vargas 2003). While the resulting utterance may be inarguably sophisticated, the impact may be not as forceful. We believe that the architecture of response generation requires more than grammar rules or some statistical distributions of language, but more intuitive in the sense that the response robustly satisfies the intention of input utterance while maintaining coherence and relevance, regardless of the surface presentation.

1.3 Objectives of research

The goal of this research is to introduce a new architecture for response generation in dialogue systems based on speaker's intentions, under statistical methodology. The new architecture is called classification-and-ranking, whereby a response utterance 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. To achieve this, the following tasks must be accomplished:

- To construct an intention-based feature model of user utterances on the basis of semantics and pragmatics interpretation of the utterances. The features will be employed for the response classification task.
- To construct an information-based attribute model of response utterances based on the semantics and informativeness of the responses. The attributes will be used to weigh response utterance during the ranking task.
- To develop a classification module. This module performs classification of user utterances into predefined response classes, based on semantics and pragmatics interpretation such that the response is coherent to the input utterance.
- To develop a ranking module. This module evaluates the semantic parity and informativeness among response utterances in a particular response class, such that they can be ranked as more relevant from one another.
- To validate the architecture through experiments using second body of dialogue corpus from different domain.

Essentially, we attempt to integrate theories from pragmatics that provide models of language, informatics that provide models of communication, and statistics that provide tools for building such models.

1.4 Scope of research

In our corpus-based approach to response generation, we limit our investigation to the type of mixed-initiative, transaction dialogues in human-machine conversation. Mixed-initiative interaction allows both dialogue participants to share equal control over the subject of conversation, while transaction dialogue is a type of task-oriented dialogue, where two dialogue participants negotiate to achieve common goal. In addition, we also assume a dialogue corpus that is readily annotated with dialogue acts. Dialogue acts are the basic building blocks for the architecture of intention-based response generation.

1.5 Research methodology

The methodology followed for the implementation and testing the classification-and-ranking architecture for response generation is summarized below. Figure 1.1 schematically illustrates the detailed methodology steps.

- Problem identification – Recognizing the importance of intentionality in response generation and reviewing statistical-based approaches to date, in effort to cater intentionality under statistical framework.