A Dialogue Analysis Model with Statistical Speech Act Processing for Dialogue Machine Translation*

Jae-won Lee and **Gil Chang Kim**
Dept. of Computer Science and CAIR
Korea Advanced Institute of Science and Technology, Taejon, 305-701, Korea
{jwonlee,gckim}@csone.kaist.ac.kr

Jungyun Seo
Dept. of Computer Science
Sogang University
Seoul, 121-742, Korea
seo@nlpeng.sogang.ac.kr

Abstract

In some cases, to make a proper translation of an utterance in a dialogue, the system needs various information about context. In this paper, we propose a statistical dialogue analysis model based on speech acts for Korean-English dialogue machine translation. The model uses syntactic patterns and N-grams reflecting the hierarchical discourse structures of dialogues. The syntactic pattern includes the syntactic features that are related with the language dependent expressions of speech acts. The N-gram of speech acts based on hierarchical recency approximates the context. Our experimental results with trigram showed that the proposed model achieved 78.59 % accuracy for the top candidate and 99.06 % for the top four candidates even though the size of the training corpus is relatively small. The proposed model can be integrated with other approaches for an efficient and robust analysis of dialogues.

1 Introduction

Recently, special concerns are paid to research on dialogue machine translation. Many different aspects of dialogue, however, make it difficult to translate spoken language with conventional machine translation techniques. One of the reasons is that a surface utterance may represent several ambiguous meanings depending on context. That means such utterance can be translated into many different ways depending on context. Interpreting this kind of utterances often requires the analysis of contexts. Therefore, the discourse structure of a dialogue plays a very important role in translating the utterances in the dialogue. Discourse structures of dialogues are usually represented as hierarchical structures which reflect embedding subdialogues (Grosz and Sidner 1986).

Many researchers have studied the way how to analyze dialogues. One of the representative approaches is the plan-based method (Litman et al. 1987; Caberry 1989). Considering that our dialogue translation system is to be combined with the speech system to develop an automatic translating telephone, however, the plan-based approach has some limitations. In an automatic translating telephone environment, the system must make one correct translated target sentence for each source sentence and must be able to respond in real time. However, the plan inference is computationally expensive and is hard to be scaled up. In order to overcome such limitations, we have focused on defining minimal approach which uses knowledgebase as small as possible while it can handle ambiguous utterances.

This paper presents an efficient discourse analysis model using statistical speech act processing for Korean-English dialogue machine translation. In this model, we suggest a probabilistic model which uses surface syntactic patterns and the N-gram of speech act reflecting the hierarchical structures of dialogues to decide the speech act of an input sentence and to maintain a discourse structure. The proposed model consists of two steps : (1) identifying the syntactic pattern of an utterance (2) calculating the plausibility for possible speech acts and discourse relations.

After presenting some motivational examples in section 2, we discuss the statistical speech act processing model to analyze discourse structure in section 3. In section 4, we describe a method to analyze dialogue structure using the proposed statistical speech act processing. We discuss experimental results for the proposed model in section 5. Finally, we draw some conclusions.

2 Motivation

Translation of dialogues often requires the analysis of contexts. That is, a surface utterance may be translated differently depending on context. In this

*This research is supported in part by the ministry of information and communication of Korea.

section, we present some motivational examples.

The word '*yey*'[1] in Korean has a number of English expression such as 'yes', 'no', 'O.K:', 'Hello', 'thanks', and so on (Jae-woong Choe 1996). When the speech act of the utterance '*yey*' is *response*, it must be translated as 'yes' or 'no'. On the other hand, when the speech act of the utterance is *accept*, it must be translated as 'O.K.'. It is even used as *greeting* or *opening* in Korean. In this case, 'Hello' is an appropriate expression in English.

The verb '*kulehsupnita*' in Korean, also, may be translated differently depending on context. *Kulehsupnita* is used to accept the previous utterance in Korean. In this case, it must be translated differently depending on context. The following dialogue examples show such cases.

Dialogue 1

A : Hankwuk hotelipnikka?
(Is it Hankwuk Hotel?)
B : Yey, *kulehsupnita.*
(Yes, *It is.*)

Dialogue 2

A : Kayksil yeyyak hasyesssupnikka?
(Did you reserve a room?)
B : Yey, *kulehsupnita.*
(Yes, *I did.*)

To differentiate such cases, a translation system must analyze the context of a dialogue. Since a dialogue has a hierarchical structure than a linear structure, the discourse structure of a dialogue must be analyzed to reflect the context in translation. There are the previous plan-based approaches for analyzing context in dialogues. Since it is very difficult to have a complete knowledge, it is not easy to find a correct analysis using such knowledge bases. In this paper, we propose a statistical dialogue analysis model based on speech acts for dialogue machine translation. Such model is weaker than the dialogue analysis model which uses many difference source .of knowledge. However, it is more efficient and robust, and easy to be scaled up. We believe that this kind of minimal approach is more appropriate for a translation system.

3 Statistical Speech Act Processing

We construct a statistical dialogue model based on speech acts as follows.

Let D denote a dialogue which consists of a sequence of n utterances, $U_1, U_2, \ldots, U_n$, and let S_i denote the speech act of U_i. With this notation, $P(U_i|U_1, U_2, \ldots, U_{i-1})$ means the probability that U_i will be uttered given a sequence of utterances $U_1, U_2, \ldots, U_{i-1}$. As shown in the equation (1), we can approximate $P(U_i|U_1, U_2, \ldots, U_{i-1})$ by the product of the sentential probability $P(U_i|S_i)$ and the contextual probability $P(S_i|S_1, S_2, \ldots, S_{i-1})$ (Nagata and Morimoto 1994). In subsequent sections, we describe the details for each probability.

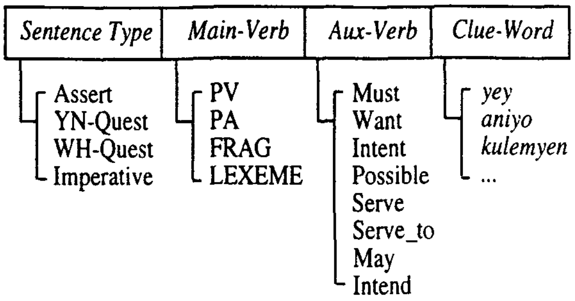

Figure 1: A Syntactic Pattern

$$
\begin{aligned}
&P(U_i|U_1, U_2, \ldots, U_{i-1}) &(1)\\
&\quad\cong\; P(U_i|S_i)P(S_i|S_1, S_2, \ldots, S_{i-1}).
\end{aligned}
$$

3.1 Sentential Probability

There is a strong relation between the speaker's speech act and the surface utterances expressing that speech act (Allen 1989 ; Andernach 1996). That is, the speaker utters a sentence which most well expresses his/her intention (speech act). This sentence allows the hearer to infer what the speaker's speech act is. However, a sentence can be used as several speech acts depending on the context of the sentence.

The sentential probability $P(U_i|S_i)$ represents the relationship between the speech acts and the features of surface sentences. In this paper, we approximate utterances with a syntactic pattern, which consists of the selected syntactic features.

We decided the syntactic pattern which consists of the fixed number of syntactic features. *Sentence Type, Main-Verb, Aux-Verb, Clue-Word* are selected as the syntactic features since they provide strong cues to infer speech acts. The features of a syntactic pattern with possible entries are shown in figure 1.

- *Sentence Type* represents the mood of an utterance. Assert, YN-Quest, WH-Quest, Imperative are possible sentence types.
- *Main-Verb* is the type of the main verb in the utterance. PA is used when the main verb represents a *state* and PV for the verbs of type

[1] All notations for Korean follow *Yale Romanization System* notation.

Table 1: A part of the syntactic patterns extracted from corpus

Speech Act	*Sentence Type*	*Main-Verb*	*Aux-Verb*	*Clue Word*
Request-Act	Imperative	PV	Request	None
Request-Act	YN-Quest	PV	Possible	None
Request-Act	Assert	PV	Want	None
Ask-Ref	WH-Quest	PV	None	None
Ask-Ref	YN-Quest	PJ	None	None
Ask-Ref	Imperative	*malhata*	Request	None
Inform	Assert	PJ	None	None
Inform	Assert	PV	None	None
Request-Conf	YN-Quest	PJ	None	None
Request-Conf	YN-Quest	FRAG	None	None
Response	Assert	PJ	None	*yey*
Suggest	Wh-Quest	PV	Serve	None

event or *action*. Utterances without verbs belong to FRAG (fragment). In the case of performative verbs (ex. promise, request, etc.), lexical items are used as a *Main-Verb* because these are closely tied with specific speech acts.

- *Aux-Verb* represents the modality such as Want, Possible, Must, and so on.
- *Clue-Word* is the special word used in the utterance having particular speech acts, such as Yes, No, O.K., and so on.

We extracted 167 pairs of speech acts and syntactic patterns from a dialogue corpus automatically using a conventional parser. As the result of applying these syntactic patterns to all utterances in corpus, we found that the average number of speech act ambiguity for each utterance is 3.07. Table 1 gives a part of the syntactic patterns extracted from corpus.

Since a syntactic pattern can be matched with several speech acts, we use sentential probability, $P(U_i|S_i)$ using the probabilistic score calculated from the corpus. Equation (2) represents the approximated sentential probability. F denotes the syntactic pattern and *freq* denotes the frequency count of its argument.

$$P(U_i|S_i) \cong P(F_i|S_i) = \frac{freq(F_i, Si)}{freq(S_i)}. \qquad (2)$$

3.2 Contextual Probability

The contextual probability $P(S_i|S_1, S_2, \ldots, S_{i-1})$ is the probability that n utterances with speech act S_i is uttered given that utterances with speech act $S_1, S_2, \ldots, S_{i-1}$ were previously uttered. Since previous speech acts constrain possible speech acts in the next utterance, contextual information have an important role in determining the speech act of an utterance. For example, if an utterance with *ask-ref* speech act uttered, then the next speech act would be one of *response*, *request-conf*, and *reject*. In this case, *response* would be the most likely candidate. The following table shows an example of the speech act bigrams.

S_{i-1}	S_i	Ratio
ask-ref	response	58.46
ask-ref	request-confirm	18.46
ask-ref	ask-if	7.69
ask-ref	ask-ref	3.08
ask-ref	suggest	3.08
ask-ref	inform	1.54

This table shows that *response* is the most likely candidate speech act of the following utterance of the utterances with *ask-ref*speech act. Also, *request-confirm* and *ask-if* are probable candidates.

Since it is impossible to consider all preceding utterances $S_1, S_2, \ldots, S_{i-1}$ as contextual information, we use the n-gram model. However, simply using n utterances linearly adjacent to an utterance as contextual information has a problem due to sub-dialogues which frequently occurred in a dialogue. Let's consider an example dialogue.

In dialogue 3, utterances 3-4 are part of an embedded segment. In utterance 3, the speaker asks for the type of rooms without responding to B's ques-

Dialogue 3

1. A : I would like to reserve a room. *request-act*
2. B : What kind of room do you want? *ask-ref*
3. A : What kind of room do you have? *ask-ref*
4. B : We have single and double rooms. *response*
5. A : A single room, please. *response*

tion (utterance 2). This subdialogue continues up to the utterance 4. As shown in the above example, dialogues cannot be viewed as a linear sequence of utterances. Rather, dialogues have a hierarchical structure. Therefore, if we use n utterances linearly adjacent to an utterance as a context, we cannot reflect the hierarchical structure of a dialogue in the model.

Therefore, we approximate the context for an utterance as speech acts of n utterances which is *hierarchically recent* to the utterance. An utterance **A** is *hierarchically recent* to an utterance **B** if **A** is adjacent to **B** in the tree structure of the discourse (Walker 1996). Equation (3) represents the approximated contextual probability in terms of hierarchical recency in the case of using trigram. In this equation, U_i is adjacent to U_j and U_j is adjacent to U_k in the discourse structure, where $1 \leq j < k \leq i-1$.

$$P(S_i|S_1, S_2, \ldots, S_{i-1}) \cong P(S_i|S_j, S_k). \tag{3}$$

4 Discourse Structure Analysis

Now we can define a discourse structure analysis model with the statistical speech act processing. Formally, choose S_i which maximizes the following probability

$$\max_{S_i} P(F_i|S_i)P(S_i|S_j, S_k). \tag{4}$$

where S_i is a possible speech act for the utterance U_i. U_j and U_k are the utterances which U_j is hierarchically adjacent to U_i, and U_k to U_j, where $1 \leq j < k \leq i-1$.

In equation (4), one problem is to search all possible U_j that U_i can be connected to. We use the dialogue transition networks (DTN) and a stack for maintaining the dialogue state efficiently. The dialogue transition networks describe possible flow of speech acts in dialogues as shown in figure 2 (Seo et al. 1994, Jin Ah Kim et al. 1995). Since DTN is defined using recursive transition networks, it can handle recursively embedded subdialogues. It works just like the RTN parser (Woods 1970). If a subdialogue is initiated, a dialogue transition network is initiated and a current state is *pushed* on the stack. On the other hand, if a subdialogue is ended, then a

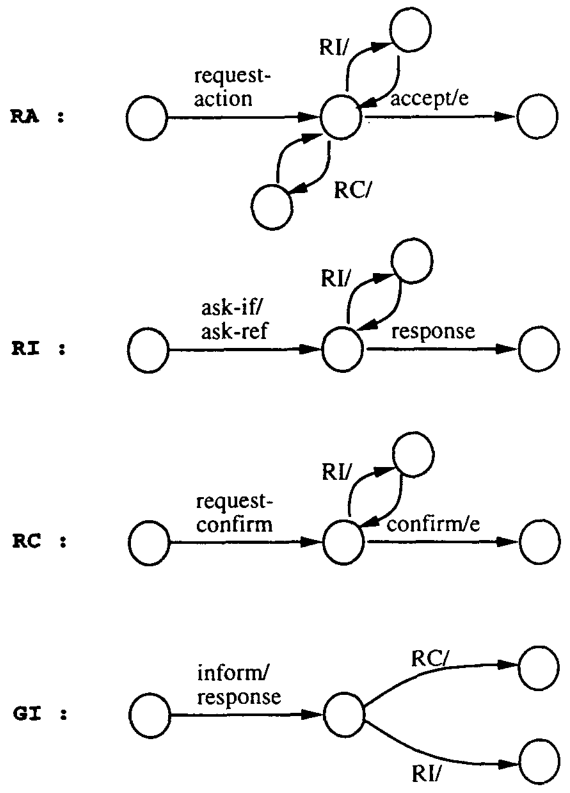

Figure 2: A part of the dialogue transition network

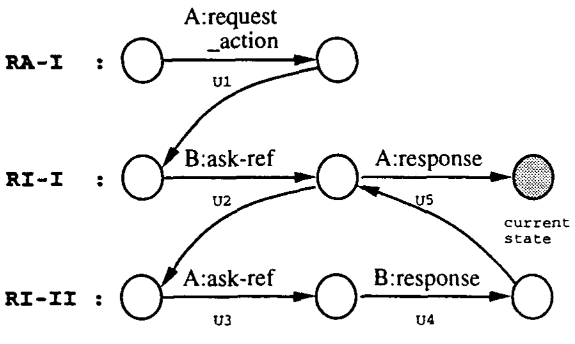

Figure 3: The transitions of dialogue 3

dialogue transition network is ended and a current state is *popped* from the stack. This process continues until a dialogue is finished.

With DTN and the stack, the system makes expectations for all possible speech acts of the next utterance. For example, let us consider dialogue 3. Figure 3 shows the transitions with the dialogue 3.

In utterance 2, according to the RA diagram in figure 2, B may *request-confirm* or *request-information*. Since B asks for the type of rooms, *push* operation occurs and a RI diagram is initiated. In utterance 3, A doesn't know the possible room sizes, hence asks B to provide such information. Therefore, *push* operation occurs again and a new RI diagram is initiated. This diagram is continued by *response* in utterance 4. In utterance 5, this diagram is popped from the stack by *response* for *ask-ref* in utterance 2.

In this state, some cases can be expected for the next utterance. The first case is to clarify the utterance 5. The second case is to return to the utterance 1. The last case is to introduce a new subdialogue. Therefore, if we assume that *ask-if* and *request-confirm* are possible from the syntactic pattern of the next utterance, then the following table can be expected for the next utterance from the dialogue transition networks.

U_k	U_j	U_i
(0:-:init)	(0:-:init)	(6:B:ask-if)
(2:B:ask-ref)	(5:A:response)	(6:B:ask-if)
(2:B:ask-ref)	(5:A:response)	(6:B:request-conf)
(0:-:init)	(1:A:request-act)	(6:B:ask-if)

Since DTN has the same expressive power as ATN(Augmented Transition Network) grammar, we believe that it is not enough to cover the whole phenomenon of dialogues. However, considering the fact that the utterances requiring context for translation is relatively small, it is practically acceptable for dialogue machine translation.

5 Experiments and Results

In order to experiment the proposed model, we used 70 dialogues recorded in real fields such as hotel reservation and airline reservation. These 70 dialogues consist of about 1,700 utterances, 8,319 words total. Each utterance in dialogues was annotated with speech acts (SA) and with discourse structure information (DS). DS is an index that represents the hierarchical structure of discourse. Table 2 shows the distribution of speech acts in this dialogue corpus. The following shows a part of an annotated dialogue corpus.

Table 2: The distribution of speech acts in corpus

Speech Act Type	*Ratio*	*Speech Act Type*	*Ratio*
ask-ref	12.30	ask-if	7.32
inform	6.35	response	19.72
request-confirm	14.65	request-action	6.84
suggest	0.20	confirm	10.64
accept	11.91	reject	0.98
correct	1.66	promise	0.59
expressive	1.86	greeting	4.10
good-bye	0.88	Total	100.00

Table 3: Experimental results

	1	2	3	4
Model I	68.48 %	74.57 %	76.09 %	76.30 %
Model II	78.59 %	92.82 %	97.88 %	99.06 %

```
/SP/hotel
/KS/Etten pangul wenhasipnikka?
/ES/What kind of room do you want?
/SA/ask-ref
/DS/[1]

/SP/customer
/KS/Etten pangi isssupnikka?
/ES/What kind of room do you have?
/SA/ask-ref
/DS/[1,1]
```

We test two models in order to verify the efficiency of the proposed model. Model-I is the proposed model based on linear recency, where an utterance U_i is always connected to the previous utterance U_{i-1}. Model-II is the model based on hierarchical recency. Table 3 shows the average accuracy of two models.

Accuracy figures shown in table 3 are computed by counting utterances that have a correct speech act and a correct discourse relation. In the closed experiments, Model-I achieved 68.48 % accuracy for the top candidate and 76.30 % for the top four candidates. In contrast, the proposed model, Model-II, achieved 78.59 % accuracy for the top candidate and 99.06 % for the top four candidates. Errors in Model-I occurred, because the hierarchical structure of dialogues was not considered. Although dialogue corpus are relatively small, the experimental results showed that the proposed model is efficient for analyzing dialogues.

6 Conclusions

In this paper, we described an efficient dialogue analysis model with statistical speech act processing. We proposed a statistical method to decide a speech act of a sentence and to maintain a discourse structure. This model uses the surface syntactic patterns of the sentence and N-gram of speech acts of the sentences which are discourse structurally recent to the sentence. Our experimental results with trigram showed that the proposed model achieved 78.59 % accuracy for the top candidate and 99.06 % for the top four candidates although the size of the training corpus is relatively small. This model is weaker than the dialogue analysis model which uses many difference source of knowledge. However, it is more efficient and robust, and easy to be scaled up. We believe that this kind of statistical approach can be integrated with other approaches for an efficient and robust analysis of dialogues.

References

Hwan Jin Choi, Young Hwan Oh, 1996, "Analysis of Intention in Spoken Dialogue Based on Learning of Intention Dependent Sentence Patterns", *Journal of Korea Science Information Society* , Vol.23, No.8, pp.862-870, In Korea.

Jae-woong Choe, 1996, "Some Issues in Conversational Analysis : Telephone Conversations for Hotel Reservation," *In Proc. of Hangul and Korean Language Information Processing*, pp.7-16, In Korea.

James F. Allen, C. Raymond Perrault, 1980, "Analyzing Intention in Utterances", *Artificial Intelligence*, Vol.15, pp.143-178

Elizabeth A. Hinkelman, James F. Allen, 1989, "Two Constraints on Speech Act Ambiguity," *In Proc. of th 27th Annual Meeting of the ACL, Association of Computational Linguistics*, pp.212-219.

Barbara J. Grosz, Candace L. Sidner, 1986, "Attention, Intentions, and the Structure of Discourse", *Computational Linguistics*, Vol.12, No.3, pp.175-204.

Philip R. Cohen, C. Raymond Perrault, 1979, "Elements of a Plan-Based Theory of Speech Acts", *Cognitive Science,* Vol.3, pp.177-212.

Diane J. Litman, James F. Allen, 1987, "A Plan Recognition Model for Subdialogues in Conversations", *Cognitive Science*, Vol.11, pp.163-200.

Hiroaki Kitano, 1994, "Speech-to-Speech Translation : A Massively Parallel Memory- Based Approach",*Kluwer Academic Publishers.*

Jan Alexandersson, Elisabeth Maier, Nobert Reithinger, 1994, "A Robust and Efficient Three-Layered Dialogue Component for a Speech-to-Speech Translation System", *Proc. of the 7th European Association for Computational Linguistics*, pp.188-193.

Jin Ah Kim, Young Hwan Cho, Jae-won Lee, Gil Chang Kim, 1995, "A Response Generation in Dialogue System based on Dialogue Flow Diagrams," *Natural Language Processing Pacific Rim Symposium*, pp.634-639.

Jungyun Seo, Jae-won Lee, Jae-Hoon Kim, Jeong-Mi Cho, Chang-Hyun Kim, and Gil Chang Kim, 1994, "Dialogue Machine Translation Using a Dialogue Model", *Proc. of China-Korea Joint Symposium on Machine Translation*, pp.55-63.

Masaaki Nagata and Tsuyoshi Morimoto, 1994, "First Steps towards Statistical Modeling of Dialogue to Predict the Speech Act Type of the Next Utterance", *Speech Communication*, Vol.15, pp.193-203.

Massko Kume, Gayle K. Sato, Kei Yoshimoto, 1990, "A Descriptive Framework for Translating Speaker's Meaning", *Proc. of the 4th European Association for Computational Linguistics*, pp.264-271.

Marilyn Walker and Steve Whittaker, 1990, "Mixed initiative in Dialogue : An Investigation into Discourse Segmentation", *In Proc. of the 28th Annual Meeting of the ACL, Association of Computational Linguistics*, pp.70-78.

Sandra Caberry, 1989, "A Pragmatics-Based Approach to Ellipsis Resolution", *Computational Linguistics*, Vol.15, No.2, pp.75-96.

Toine Andernach, 1996, "A Machine Learning Approach to the Classification of Dialogue Utterances", Proceedings of NeMLaP-2, Bilkent University, Turkey.

Marilyn A. Walker, 1996, "Limited Attention and Discourse Structure,", *Computational Linguistics*, Vol.22, No.2, pp.255-264.

Woods, W. A., 1970, "Transition Network Grammars for Natural Language Analysis," *Commun. of the ACM*, Vol.13, pp.591-606.