Vol. 35 No. 6 Transactions of Information Processing Society of Japan June 1994

Regular Paper

An Information-Theoretic Model of Discourse for Next Utterance Type Prediction

MASAAKI NAGATA [†,*] and TSUYOSHI MORIMOTO [†,**]

We propose a statistical model of dialogue that is based on an information-theoretic interpretation of discourse, to predict the illocutionary force type of the next utterance. The model consists of a second-order Markov model of utterances classified by their illocutionary force types, such as REQUEST, INFORM, etc., and it gives us a criterion for measuring whether the speech recognition candidate forms a natural local discourse in terms of the speech act sequence. By predicting the next utterance in an abstract level, we can rule out erroneous speech recognition candidates that are syntactically and semantically correct, but contextually incorrect. We show the effectiveness of the statistical dialogue model for utterance type prediction by extensive experiments using 100 telephone dialogues containing 7,531 utterances. The model achieves 61.7 % accuracy for the top candidate and 85.1 % for the top three candidates, when 50 dialogues were used for training and the other 50 dialogues were used for testing. We also show that the model can capture the basic characteristics of the local discourse structure, such as turn-taking and speech act sequencing, and dialogue-type dependent features, such as initiative, which is the allocation of the control and the manner by which the control is transferred.

1. Introduction

Recent progress in speech recognition technology has introduced a new set of problems to natural language processing. One exciting challenge is the use of pragmatic information in speech recognition. Some erroneous recognition candidates are hard to rule out because they are syntactically and semantically correct as sentences but do not fit the context. Discourse, or context information, should play an important role in solving this problem. However, it is not yet clear what kind of contextual constraints are effective in speech recognition.

There are several pioneering works that try to use pragmatic information in speech recognition. MINDS[5] uses knowledge of dialog structures, user goals and focus in a problem solving situation, and dynamically generates semantic network grammars. SOUL[16] uses an extensive semantic and pragmatic knowledge base for abductive reasoning and constraint satisfaction in order to deal with ambiguous, illegal, and context removable utterances. Yamaoka[15] proposed a dialogue interpreting model based on the plan recognition technique and used it for predicting the appropriate type of following utterance.

One bottleneck in these plan-based approaches is that they have to hand-code the dialog structures and higher level knowledge into the system. It would be a painful task to provide this knowledge for larger domains and even larger vocabularies. It is also to be noted that these plan-based approaches can not prioritize their predictions, which is definitely needed for practical applications, such as speech recognition.

We propose a statistical model of discourse which is based on the information-theoretic interpretation of a dialogue. This model is an alternative to the previous plan-based approaches to next utterance prediction applications. The dialogue model consists of a second-order Markov model (trigram) of utterances classified by their Illocutionary Force Types

† ATR Interpreting Telephony Research Laboratories
* Presently with NTT Network Information Systems Laboratories
** Presently with ATR Interpreting Telecommunications Research Laboratories

Table 1 Most frequently found speech recognition errors in the model dialogues.

Uttered	Recognized
[hai]:193(8.4%)	[na-i]:43, [haya-i]:39, [a-ru]:22, [chika-i]:19, [haya-i]:19, [atsu-i]:18, [a-i-ta]:14, [atsu-i]:13, [taka-i]:3, [ki-tai]:3
desu]:181(7.9%)	desu-ka]:101, deshi-ta]:45, desu-Qke]:19, desu-ga]:11, zero-de-iraQshai-masu-ka]:3, desu-ne]:2
masu]:164(7.1%)	masu-ka]:115, masi-ta]:19, masu-ga]:19, ru-n-desu]:3, mase-n]:2, masu-ne]:2, ta]:1, te-a-Qta]:1, ta-N-desu]:1, te-shima-u-ka]:1
ta]:109(4.7%)	ta-ka]:50, ta-ga]:50, ta-ne]:9
o]:104(4.5%)	mo]:87, to]:10, no]:5, ha]:1, ni-mo]:1
ϕ:85(3.7%)	e:24, ni:18, de:10, rare:6, N:5, nado:4, san:4, na-N:2, te:2, na-no:2, nee:1, gatsu:1, de]-[to-Q-te-i-ru:1, ni:1, go:1, ita:1, to:1, dake:1
wa]:77(3.3%)	ga]:70, ya]:3, o]:3, mo]:1
ga]:76(3.3%)	ka]:47, yara]:8, wa]:8, ya]:5, ni-wa]:4, ne]:2, de-wa]:2
desu-ka]:59(2.6%)	deshi-ta]:53, deshi-ta-ga]:4, N-deshi-ta]:2
no]:52(2.3%)	mo]:33, o]:6, nado-o]:5, to]:3, ni-mo]:2, ga]:2, hodo]:1

(IFTs). IFT[8] is an abstraction of the speaker's intention in terms of the type of action the speaker intends by the utterance. This dialogue model is designed to serve as a knowledge source of turn-taking and speech act sequencing, and can be used for ruling out incoherent erroneous recognition candidates by predicting the next utterance in an abstract level.

The advantage of the statistical dialogue model is that it does not require hand-crafted plans, which are descriptions of possible action sequences in the domain. It can learn its parameters from the training examples, namely, an annotated dialogue corpus. Another important advantage of the statistical model is that it is easy to combine statistical information about the discourse structure with other statistical information, such as speech recognition scores and word frequencies, in a uniform fashion.

In the following sections, we first describe the motivation behind the statistical dialogue model, and explain the model based on the speech act theory and the information theory. We then report the experimental results to prove the effectiveness of the dialogue model, and discuss its relevance to the previous research on discourse.

2. Needs for Discourse in SR

Table 1 shows the 10 most frequently found speech recognition errors contained in the top 3 candidates of ATR's speech recognizer,[7] when tested on 4 speakers (2 males and 2 females) using the "model dialogues," which will be described later.* The inputs ("Uttered") and outputs ("Recognized") are associated with the number of occurrences in which that type of recognition error happened.**

In general, we found that many speech recognition errors involve sentence final particles, tuch as *ka* and *ga*. We also found that fixed short expressions, such as *hai, iie,* and *moshimoshi*, are often confused with simple short assertions. These kinds of errors are hard to rule out

* In Table 1, "[" and "]" indicate the beginning and the end of a *bunsetsu* phrase. "-" indicates the concatenation of morphemes. ϕ indicates an empty string, so ϕ in the input indicates an addition, while ϕ in the output indicates deletion.

** We treat homonymous morphemes separately when they have different orthography in Chinese characters, such as *atsu-i* (厚い/thick) and *atsu-i* (暑い/hot).

Secretary: *tourokuyoushi-wa sudeni omochi-desho-u-ka*
(Do you already have a registration form?)
Questioner: *iie mada-desu*
(No. Not, yet.)
Secretary: *waka-ri-mashi-ta*
(I see.)

Incorrect Recognition Examples;
> *a-ri-mashi-ta*
([(Here) it] is.)
> *na-ri-mashi-ta*
([it] becomes [something].)

Fig. 1 Incorrect recognition examples for fixed short expressions.

because they are syntactically and semantically correct as sentences. However, they could be ruled out if we can use information about the local discourse structure, such as turn-taking and proper speech act sequences.

For example, **Fig. 1** shows typical speech recognition results of the Japanese confirmation expression "*waka-ri-mashi-ta*" in context. The expression is often misrecognized as several different sentences, such as "*a-ri-mashi-ta, "na-ri-mashi-ta*" and "*kaka-ri-mashi-ta,"* all of which are assertions rather than confirmations. This is mainly caused by the difficulty in recognizing the semi-vowel "w" at the beginning of the utterance. Since the last utterance is issued within a typical discourse structure of a question-answer-confirmation cycle, it is reasonable to give a positive bias to those recognition candidates that define the utterance as being a confirmation.

Figure 2 shows typical speech recognition results of the Japanese sentence final expression "*desu-ga*" in context. The expression consists of the copula "*desu*" followed by the softener "*ga,*" and it is often mistaken for "*desu-ka,*" which is the copula followed by the question marker "*ka.*" This recognition error results in ambiguity between the declarative and interrogative sentences. However, it is inevitable because the input speech signal is relatively weak in the final part of an utterance, and it is very difficult to distinguish the voiced consonant "g" from its unvoiced counterpart "k." In this interaction, the secretary asks the questioner whether there is somethings that he can do for him. Therefore, an utterance representing the questioner's request is expected as a response. Of course, the questioner is not prohibited from asking a question in response to the first question, but, in this corpus and in this context, it seems less probable

Secretary: *dono-youna goyoukeN-desho-u-ka*
(What can I do for you?)
Questioner: *kaingi-ni moushiko-mi-tai-no-desu-ga*
(I would like to register for the conference.)

Incorrect Recognition Examples;
> *kaingi-ni moushiko-mi-tai-N-desu-ga*
(I would like to register for the conference.)
> *kaingi-ni moushiko-mi-tai-N-desu-ka*
(Do you want to register for the conference?)

Fig. 2 Incorrect recognition examples for sentence final modal expressions.

that the questioner would ask a question than make a request.

In general, short answer and modal expressions are inherently context-dependent and it is virtually impossible to measure their linguistic appropriateness without context. Therefore, we conclude that a dialogue model that can assess the linguistic plausibility of sentential candidates in the context is definitely needed if we are to improve the speech recognition accuracy for dialogues.

3. Illocutionary Force Types

3.1 Classification of Utterances

When the speaker utters a sentence, the hearer receives communicative signs in addition to the propositional content. According to the speech act theory, these signs are classified as illocutionary forces governed by certain felicity conditions.[1),13)] We have classified surface Illocutionary Force Types (IFTs) into 9 types,[8)] as listed in **Table 2**. We have adopted surface IFTs as the primitive with which the knowledge of interaction patterns are represented, because surface IFTs have a relatively straightforward correspondence to the surface syntactic patterns.

Phatics are phatic expessions, such as the greetings "hello" and "good-bye" that appear at the opening and closing of a dialogue. **Expressive**s are idioms that express the speaker's feeling, such as "thank you" and "you're welcome." **Responses** are idiomatic responses and short answers. In **promise**s, the speaker commits himself to perform an action, while in requests, the speaker asks the hearer to perform an action. **Questionif**s are Yes-No questions and **questionref**s are WH questions. **Questionconf**s are confirmations. Moreover, we have introduced a special utterance type **DBM** (Discourse Boundary Marker) which stands for the beginning and the end of a dialogue.

Table 2 Surface illocutionary force types.

Surface IFT	Examples
phatic	*moshimoshi* (Hello) *shiturei-shi-masu* (Good-bye)
expressive	*arigatou-gozaima-shi-ta* (Thank you) *yoroshiku onegai-shi-masu* (Thank you)
response	*hai* (Yes) *sou-desu* (That's right) *waka-ri-mashi-ta* (I see)
promise	*tourokuyoushi-o oku-ra-se-te-itada-ki-masu* (I will send you a registration form.)
request	*chikatetsu-de kitaoozieki-made i-Q-te-kudasai* (Please go to Kitaoozi station by subway.)
inform	*koNkai-wa waribiki-o okona-Q-te-ori-mase-N* (We are not giving any discount this time.)
questionif	*kaigi-no aNnaisho-wa omochi-desu-ka* (Do you have the announcement of the conference?)
questionref	*dou-sure-ba yorosii-desu-ka* (What should I do?)
questionconf	*sudeni tourokuryou-o* *huriko-ma-re-te-ora-re-masu-ne* (You have already transferred the registration fee, right?)

Discourse structure can be classified into local and global. The local discourse structure refers to linguistic constraints between adjacent and neighboring utterances, whereas the global discourse structure refers to the overall organization of a conversation, such as the opening section, the closing section, and the embedding of discourse segments. Although our dialogue model is mainly designed for the local discourse structure, the opening and the closing section are treated in the framework by introducing a special utterance type, **DBM**, that indicates the beginning and the end of a conversation, as well as providing specific utterance types for idiomatic expressions, **phatic** and **expressive**, that appear in openings and closings.

3.2 Identification of IFTs

In order to identify the IFT of the utterance, we have implemented a descriptive framework for speech act inference[8] using a typed feature structure rewriting system[4] as an inference engine.* It accepts the typed feature structure of

* Although the IFT identification module currently covers only the 10 model dialogues, we believe the framework can be extended for general use.

```
Utterance: tourokuyoushi-o shikyuu oku-ra-se-te-itada-ki-masu
 (''I will send you the registration form immediately.'')

<<Input of IFT Analysis>>
[[reln  temorau-receive_favor]
 [aspt  unrl]]
 [agen ?x03[[label *speaker*]]]
 [recp ?x04[[label *hearer*]]]
 [obje [[reln  saseru-permissive]
        [agen ?x04]
        [recp ?x03[]]
        [obje [[reln  okuru-1]
               [agen ?x03]
               [recp ?x04]
               [mann [[parm ?x02[]]
                      [restr [[reln  sikyuu-1]
                              [entity ?x02]]]]]
               [obje [[parm ?x01[]]
                      [restr [[reln  tourokuyousi-1]
                              [entity ?x01]]]]]]]]]]
<<Output of IFT Analysis>>
[[reln  promise]
 [aspt  unrl]]
 [agen ?x03[[label *speaker*]]]
 [recp ?x04[[label *hearer*]]]
 [obje [[reln  okuru-1]
        [agen ?x03]
        [recp ?x04]
        [mann [[parm ?x02[]]
               [restr [[reln  sikyuu-1]
                       [entity ?x02]]]]]
        [obje [[parm ?x01[]]
               [restr [[reln  tourokuyousi-1]
                       [entity ?x01]]]]]]]]]]
```

Fig. 3 An example of IFT inference.

the input utterance, which is output by the HPSG parser[11] as a semantic representation. It then identifies the IFT using IFT inference rules, and outputs a typed feature structure whose outermost relation name is the resulting IFT.

Figure 3 shows an example of IFT inference. IFT inference rules are basically a set of pattern-action rules that capture a certain pattern in the semantic representation and rewrite it according to the rule. For example, if the rewriting system receives the feature structure of the utterance "*tourokuyoushi-o shikyuu oku-ra-se-te-itada-ki-masu*," which means "I will send you the registration form immediately," it detects the feature structure corresponding to "*se-te-itada-ki-masu*,"* and rewrites it to the feature structure whose IFT is **promise**, by applying the following IFT inference rule.

saseru-permissive + *temorau*-receive_favor
⇒**promise** (1)

4. Information-Theoretic Model of Discourse

4.1 Statistical Modeling of Dialogue

We consider human dialogue to be an information source which outputs symbols chosen from some finite set of vocabulary. Although we cannot enumerate the variations of human utterances, we approximate them by a finite set of symbols, using speech act types (IFTs) as a codebook for utterances. This information-theoretic interpretation of dialogue leads us to a statistical dialogue model as follows.

Let $\mathbf{S}$ denote a sequence of n sentences that constitute a dialogue.

$$\mathbf{S}=S_1, S_2, \cdots, S_n \quad (2)$$

We approximate $\mathbf{S}$ by $\mathbf{s}$, which denotes a sequence of n speech act types (IFTs),**

$$\mathbf{s}=s_1, s_2, \cdots, s_n \quad (3)$$

where s_i is the speech act type of S_i. Hence, we will think of $P(\mathbf{s})$ as a statistical model of dialogue. Using elementary rules of probability theory, the dialogue model's probability $P(\mathbf{s})$ can be formally decomposed as

$$P(\mathbf{s})=\prod_{i=1}^{n} P(s_i|s_1, \cdots, s_{i-1}) \quad (4)$$

where $P(s_i|s_1, \cdots, s_{i-1})$ is the probability that a sentence with speech act type s_i is uttered, given that sentences with speech act types $s_1, \cdots, s_{i-1}$ were previously uttered.

In practice, the probability $P(s_i|s_1, \cdots, s_{i-1})$ would be very difficult to estimate since most histories $s_1, \cdots, s_{i-1}$ would have occurred only a very few times for larger values of i. Therefore, we approximate it by the trigram probabilities.

$$P(\mathbf{s})=\prod_{i=1}^{n} P(s_i|s_{i-2}, s_{i-1}) \quad (5)$$

Originally we tried to estimate the trigram probabilities $P(s_i|s_{i-2}, s_{i-1})$ by the relative frequency $f(s_i|s_{i-2}, s_{i-1})$

$$P(s_i|s_{i-2}, s_{i-1})=f(s_i|s_{i-2}, s_{i-1})=\frac{C(s_{i-2}, s_{i-1}, s_i)}{C(s_{i-2}, s_{i-1})} \quad (6)$$

where the function C counts the number of speech act type sequences in its argument. Since, at the moment, we do not have a large dialogue corpus annotated with speech act types, the above estimation is inadequate. In fact, when we divided a corpus of 100 telephone dialogues. (7,631 sentences) into test and training subsets (50 dialogues each), and when we use 19 symbols (Speaker-IFT pairs) as sentence tags, we found that 40 % of the trigrams appearing in the test subset never appeared in the training subset, so it is necessary to smooth the trigram frequencies.

We employed the standard smoothing technique known as "deleted interpolation," which simply interpolates trigram, bigram, unigram, and zerogram relative frequencies,[6]

$$P(s_i|s_{i-2}, s_{i-1})=q_3 f(s_i|s_{i-2}, s_{i-1})+q_2 f(s_i|s_{i-1})+q_1 f(s_i)+q_0 V(s) \quad (7)$$

where the nonnegative weights satisfy $q_3+q_2+q_1+q_0=1$. The weights q_i are chosen to satisfy the maximum-likelihood criterion, that is, they are adjusted to maximize the probability of the observed data.

4.2 Quality Measures for Dialogue Model

Now we can define an objective measure of dialogue model quality. According to the information theory, the entropy of the m-th order Markov source can be given from the information source alphabet S and the conditional probabilities.

* "*se-ru*" is a causative auxiliary verb. "*itada-ki-masu*" is a polite form of "*mora-u*", which is a benefactive auxiliary verb whose meaning shows that the speaker receives a favor.

** In fact, s_i can be anything that quantifies the discourse-related aspects of the sentence S_i.

$$H(S)=-\sum_{s^{m+1}} P(s_{j1}, s_{j2}, \cdots, s_{jm}, s_i) \times \log P(s_i|s_{j1}, s_{j2}, \cdots, s_{jm}) \quad (8)$$

If you think of dialogue as an information source whose output symbol is the combination of the speaker and the IFT of an utterance, you can define *discourse entropy* per sentence with respect to IFT, using Eq. (8). We can also define *discourse perplexity* per sentence with respect to IFT as follows.

$$DP=2^{H(S)} \quad (9)$$

To the first order of approximation, which disregards the difference of difficulties in recognizing, parsing and identifying IFTs of utterances, the next utterance type prediction task can be considered as difficult as the prediction of a dialogue with *DP* equally likely utterance types. Discourse perplexity is therefore a measure of the average branching of the dialogue with respect to IFT when the dialogue model is presented.

We can also define an objective measure for the significance of adjacent speech act pairs, using mutual information. Mutual information, $I(x\,;y)$, compares the probability of observing speech act type x and speech act type y, together with the probabilities of observing x and y independently.

$$I(x\,;y)=\log_2\frac{P(x, y)}{P(x)P(y)} \quad (10)$$

If there is a genuine association between x and y, then $I(x\,;y)\gg 0$. If there is no relationship between x and y, then $I(x\,;y)\approx 0$. If x and y are in complementary distribution, then $I(x\,;y)\ll 0$. $P(x)$ and $P(y)$ are estimated by counting the number of observation of x and y in a corpus, $f(x)$ and $f(y)$, and normalizing by N, the size of the corpus. Joint Probabilities, $P(x, y)$, are estimated by counting the number of times that x is followed by y in a window of w utterances, $f_w(x, y)$, and normalizing by $N(w-1)$.

5. Experiment

5.1 Dialogue Corpora and their Entropy

For evaluating the statistical dialogue model, we manually annotated the IFT of each utterance for three different kind of corpora. The task domain is registration of an international conference in which a conversation between a secretary and a questioner is carried out by telephone or keyboard.

The first corpus is what we call "Model Dialogues," which consists of 10 dialogues with 225 sentences. They are a collection of the scripts of typical conversations that would appear in the conference registration task. Their topic includes registering the conference, reserving a room in a hotel, inquiring about a sightseeing tour, etc.. They are used as a benchmark for SL-TRANS,[10] ATR's speech-to-speech translation system. The second corpus is "Telephone Dialogues," which consists of 100 dialogues with 7,531 sentences. The third corpus is "Keyboard Dialogues," which consists of 50 dialogues with 1,686 sentences. The second and the third sets of dialogues are taken from the ATR Dialogue Database (ADD),[2] which collects simulated telephone or keyboard dialogues spontaneously spoken or typed in Japanese or English.

For annotating the corpora, we made an IFT annotation manual. The IFT annotation criteria of the manual is, basically, an extension of the IFT inference rules used in SL-TRANS, as described in Section 3.2, that covers the 10 model dialogues. We then trained an annotator, who experienced the development of a lexicon of machine translation systems, and had her manually label the IFT of each utterance. It took about two months to annotate 100 telephone dialogues and 50 keyboard dialogues, including the time taken for deciding the specific rules for annotation and checking data consistency.

Table 3 shows the distribution of IFTs for the three corpora. You can see that **inform** is the utterance type that most frequently occurred in all dialogues. One of the major differences between telephone and keyboard dialogues is that telephone dialogues have more **response**

Table 3 Distribution of IFTs (percent).

IFT	Model	Telephone	Keyboard
expressive	4.4	3.4	6.9
inform	39.6	36.9	41.0
request	11.1	5.7	12.6
response	23.6	34.6	11.7
questionref	6.7	3.0	8.2
phatic	4.4	4.8	3.0
questionif	6.7	8.6	13.1
promise	0.9	1.1	1.2
questionconf	2.7	2.0	2.1

Table 4 The discourse entropy with respect to IFT.

	Model Dialogues		Telephone	Keyboard
	IFT	SP-IFT	SP-IFT	SP-IFT
zerogram	3.32	4.25	4.25	4.25
unigram	2.64	3.50	3.29	3.57
bigram	2.04	2.15	2.48	2.95
trigram	1.45	1.26	2.02	2.19

utterances than keyboard dialogues, which probably reflects the bandwidth available for communication. Model dialogues can be characterized as being intermediate between the telephone and keyboard dialogues.

Table 4 shows the discourse entropies of dialogue models of different history lengths, namely, zerogram, unigram, bigram, and trigram for the three corpora. For model dialogues, it also shows the discourse entropies of two different utterance encoding methods: The first method uses the simple IFT of the utterance, and the second method uses the combination of the speaker and the IFT (SP-IFT).*

Table 4 shows that a drastic reduction in discourse entropy with respect to IFT is achieved by using trigrams of utterance type sequences with the speaker label (SP-IFT). It also shows that the reduction of the discourse entropy by the SP-IFT trigram holds regardless of the modality of communication, telephone or keyboard. We assume this property results from the basic nature of dialogues, that is, two adjacent utterances are never independent. For example, if one party asks a question, the other party tries to answer it.

Note that when SP-IFT is used to code an utterance, the discourse entropy measure includes the effect of the nondeterminism of the next speaker. In other words, the total number of symbols in the codebook is not 10 but 19. For example, as the discourse entropy for the model dialogues using SP-IFT trigram is 1.26, its discourse perplexity is 2.3. This means that 19 symbols are reduced to 2.3 equally likely symbols with the SP-IFT trigram, whereas, 10 symbols are reduced to 2.7 equally likely symbols with the IFT trigram.

5.2 IFT Ambiguity in Speech Recognition

In order to show the necessity for the dialogue model, we first investigated[12] how much ambiguity the speech rocognition results have with respect to IFT. The speech material was the 10 model dialogues uttered by one male speaker. We examined the top 10 sentential candidates as output by the speech recognizer.[7] Among the 225 utterances in the model dialogues, 213 utterances have some kind of sentential outputs that are recognized as syntactically and semantically correct by the HPSG parser.[11] There are 98 utterances that have IFT ambiguity, with 83 utterances having two kind of IFTs in the recognition candidates and 15 utterances having three. The average IFT ambiguity is 1.53. When we used the model dialogues for both training and testing, we found that the IFT trigram predicted the correct IFT in more than 70 % of the utterances exhibiting IFT ambiguity.

5.3 IFT Prediction Accuracy

In order to prove the effectiveness of the statistical dialogue model, we tested the accuracy of predicting the IFT of the next utterance using the SP-IFT trigram and the two previous utterances. In the following experiments, if not explicitly mentioned, the statistical dialogue model is trained and tested using the same corpus. However, since the parameters are smoothed as mentioned in Section 4.1, the model has already been generalized to some extent. This was shown by the comparison of open and closed experiments.

An intuitive feeling for the quality of the statistical dialogue model may be obtained from **Fig. 4**, which shows the top-three IFTs that are predicted to be most likely, given the two preceding speakers and IFTs, as well as the SP-IFT trigram. For example, at the 10th sentence in the first model dialogue d01, knowing the two preceding speakers and IFTs, i.e. **Secretary**'s **request** and **Secretary**'s **questionif**, the dialogue model predicts that the most likely next IFT is **response**, with a probability of 0.630361, if the next speaker is **Questioner**.

Figure 5 shows the IFT prediction accuracy for the three corpora. If we examine the top three candidates, we get IFT prediction accuracies of 93.3 %, 89.1 %, and 83.9 % for the model, telephone, and keyboard dialogues, respectively. Considering the IFT distribution shown in Table 3, these prediction accuracies are significantly higher than the accumulated

* From now on, we will use SP-IFT to represent the combination of the speaker and the IFT.

```
d01 5 S-questionref dono-youna goyoukeN-desyo-u-ka
                    (''May I help you?'')
                    (lit. ''What kind of business do you have?'')
S-response S-response S-inform 0.525042
S-response S-response S-questionref 0.108865
S-response S-response S-response 0.070138

d01 6 Q-inform kaigi-ni moushiko-mi-tai-no-desu-ga
                (''I'd like to apply for the conference.'')
S-response S-questionref Q-inform 0.762553
S-response S-questionref Q-response 0.021947
S-response S-questionref Q-questionref 0.013192

d01 7 Q-questionref dono-youna tetsuzuki-o sure-ba yoroshi-i-no-desyo-u-ka
                    (''What kind of procedure should I go through?'')
S-questionref Q-inform Q-inform 0.206922
S-questionref Q-inform Q-questionref 0.132455
S-questionref Q-inform Q-questionif 0.098266

d01 8 S-request tourokuyoushi-de tetsuzuki-o shi-te-kudasai
                (''Please proceed by using the registration form.'')
Q-inform Q-questionref S-inform 0.464212
Q-inform Q-questionref S-request 0.241136
Q-inform Q-questionref S-response 0.104383

d01 9 S-questionif tourokuyoushi-wa sudeni omochi-desyo-u-ka
                  (''Do you already have the registration form?'')
Q-questionref S-request S-inform 0.202862
Q-questionref S-request S-request 0.151456
Q-questionref S-request S-questionif 0.077100

d01 10 Q-response iie
                 (''No.'')
S-request S-questionif Q-response 0.630361
S-request S-questionif Q-questionref 0.139912
S-request S-questionif Q-inform 0.027419

d01 11 Q-inform mada-desu
                (''Not yet.'')
S-questionif Q-response Q-inform 0.316685
S-questionif Q-response Q-expressive 0.128883
S-questionif Q-response Q-response 0.115814

d01 12 S-response waka-ri-mashi-ta
                  (''I see.'')
Q-response Q-inform S-response 0.149529
Q-response Q-inform S-questionconf 0.104336
Q-response Q-inform S-request 0.081634

d01 13 S-inform sore-de-wa tourokuyoushi-o ookuri-ita-shi-masu
                (''Then, I'll send you the registration form.'')
Q-inform S-response S-inform 0.541964
Q-inform S-response S-questionref 0.075021
Q-inform S-response S-response 0.070138
```

Fig. 4 IFT prediction examples (model dialogues).

sum of the top three IFTs, namely, 74.3 %, 80.1 %, and 66.7 %, for the three corpora, respectively.

In general, model dialogues achieve the highest IFT prediction accuracy. Telephone dialogues and keyboard dialogues follow in decreasing order. This reflects the values of the discourse entropies shown in Table 4. However, only for the first candidate, the prediction accuracy of the model dialogues is lower than that of the telephone dialogues. This is caused by insufficient training data. In other words, the smoothing significantly degrades the prediction accuracy of the top candidate in the case of the model dialogues. If we do not apply parameter smoothing, the IFT prediction accuracy of the top candidate for the model dialogues is 82.2 %, which is higher than that of the telephone dialogues.

Table 5 IFT prediction accuracy (telephone).

Rank	Closed	Open
1	65.6	61.7
2	82.2	77.5
3	90.1	85.1
4	94.4	90.0
5	97.0	93.6

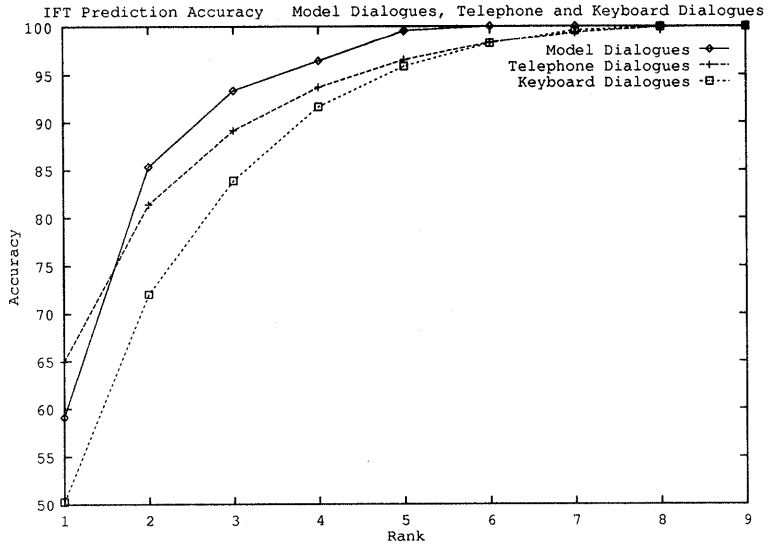


Fig. 5 IFT prediction accuracy.

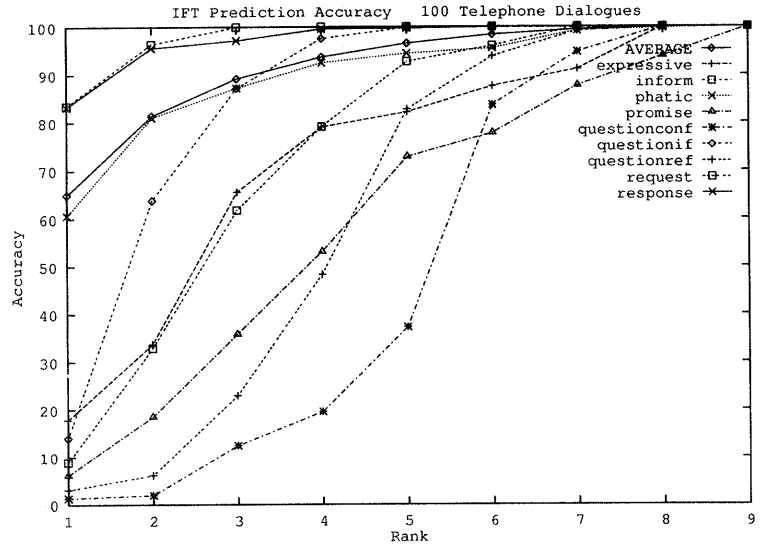


Fig. 6 IFT prediction accuracy with respect to IFT.

To show the generality of the statistical model, we tested it on open data. We divided the 100 telephone dialogues into two subsets of 50 dialogues, one for training and the other for testing. **Table 5** shows the average IFT prediction accuracy of the open and closed experiments on the telephone dialogues.

In general, the prediction accuracy of the open experiments is about 5 % lower than that of the closed experiments. However, the result of the open experiments is significantly better than the accumulated sum of the N-most frequent IFTs shown in Table 3. For example, it is about 25 % higher for the top candidate, and about 5 % higher for the top three candidates.

5.4 Significance of Speech Act Pairs

Figure 6 shows the IFT prediction accuracy for the 100 telephone dialogues with respect to

Table 6 Statistically significant SP-IFT associations.

$I(x;y)$	$F(x,y)$	x	$F(x)$	y	$F(y)$
4.766883	16	S-phatic	46	DBM	100
4.226315	11	DBM	100	S-phatic	46
4.045823	23	S-expressive	109	DBM	100
3.804929	29	Q-expressive	149	S-expressive	109
3.227517	6	S-expressive	109	S-phatic	46
3.191571	8	Q-expressive	149	S-phatic	46
3.134326	35	Q-phatic	312	DBM	100
2.825777	13	S-phatic	46	Q-phatic	312
2.776533	6	S-phatic	46	Q-expressive	149
2.771716	13	Q-expressive	149	DBM	100

Table 7 Statistically significant SP-IFT associations involving questions.

$I(x;y)$	$F(x,y)$	x	$F(x)$	y	$F(y)$
2.058936	6	S-questionref	51	Q-request	221
1.785651	11	S-questionif	113	Q-request	221
1.779707	113	Q-questionref	175	S-inform	1472
1.558111	25	S-questionref	51	Q-inform	1303
1.445634	272	Q-questionif	531	S-inform	1472

each IFT. We can see that the statistical dialogue model is quite good at predicting IFTs such as **response** and **inform**, but that it is uncertain about IFTs such as **quentionif** and **questionref**. We assume that this property is strongly related to the initiative of the dialogue, which we will discuss later.

Table 6 shows the top-10 SP-IFT pairs, which have the largest amount of mutual information where the window size is 2, and which are observed more than 5 times. **Table 7** shows the top-5 SP-IFT following questions (**questionif** and **questionref**), which have the largest amount of mutual information where the window size is 2, and which are observed more than 5 times.

These tables show that speech act type pairs which appear in the opening and the closing sections of a dialogue, such as **phatic, eypressive**, and **DBM**, have the largest amount of mutual information, and that speech act pairs involving **questionif** and **questionref** have the second largest amount of mutual information.

Since there are varying degrees of significance among adjacent speech act pairs, the statistical dialogue model will not effectively predict the next utterance type at every point in the discourse. However, it is expected to yield effective predictions in certain contexts, where the mutual information of the speech act types is significantly large.

6. Discussion

Walker[14] proposed a set of rules for the allocation and transfer of control in mixed-initiative dialogue, and concluded that initiative plays an important role in the structuring of discourse. Her rules place a segment boundary whenever the "controller" is changed, and she validated this by exploring the distribution of anaphoric expressions. We assume the differences in the IFT prediction accuracy with respect to IFT shown in Fig. 6 supports her arguments in a different way. We can conclude that the utterance types of "noncontroller," such as **response** and **inform** (in her words, "prompt" and "assertion") are easy to predict, while the utterance types of controller, such as **questionif**, **questionref** and **request** (in her words, "command" and "question") are difficult to predict.

Since there is a varying degree of predictability and association between adjacent utterances, discourse segment boundaries could be placed by mutual information statistics, such as Magerman's method[9] for sentence parsing. Once we bracket a dialogue, we might be able to use the hierarchical discourse structure for identifying the referent of the referring expres-

sions, although we might not be able to determine the relationship between discourse segments, which is elaborated by Grosz.[3)]

Walker[14)] also discovered that both allocation of control and the manner in which control is transferred is strongly depends on the dialogue type. She reported that the expert has control for around 90 % of the utterances in task-oriented dialogues, whereas control is shared almost equally in advice-giving dialogues. We assume that the reduction in discourse entropy that occurs upon adding speaker information for utterance encoding reflects the bias of the allocation and transfer of control in the conference registration domain, which, we think, lies between task-oriented and advice-giving dialogues.

The classification of utterance types used for the dialogue model is open to further discussion. The current IFT uses syntactic features such as declarative and interrogative, as well as semantic features related to intention and desire. We plan to revise the current IFT to give added consideration to syntactic features such as the grammatical person of the subject/object, polarity, and semantic features such as permission, possibility, and benefit. However, it must be noted that as the number of utterance classifications increases, the amount of training data required also increases.

We found that the amount of data used for statistical modeling is still insufficient. For example, the values of q_0, q_1, q_2, q_3 in the smoothing Eq. (7) for the 100 telephone dialogues are 0.034406, 0.015718, 0.347588, 0.602289, respectively. This means, to the smoothed trigram probability, the raw trigram contributed 60 % and the raw bigram contributed 35 %. Since it is desirable for the raw trigram to contribute more than 90 % to the smoothed trigram probability, we assume that the amount of annotated dialogue data required for reliable estimation is one order of magnitude larger, say at least over 500 dialogues.

7. Conclusion

In this paper, we proposed a statistical model of dialogue for predicting the illocutionary force type of the next utterance. This model consists of a second-order Markov model of the sequence of the utterance type and its speaker. Since it is based on an information-theoretic interpretation of discourse, we can define discourse entropy as an objective measure of the quality of the dialogue model. We have shown the effectiveness of the statistical dialogue model for the utterance type prediction task by extensive experiments using 100 telephone dialogues and 50 keyboard dialogues. We have also shown that the model can capture the basic characteristics of the local discourse structure, such as turn-taking and speech act sequencing, and the dialogue-type dependent features, such as initiative, which involves the allocation of the control and the manner by which the control is transferred.

Other than the N-gram model used in this paper, there are several other possibilities for statistical modeling of dialogue. One prospective choice is the Hidden Markov Model (HMM) because it is directly equivalent to a stochastic regular grammar, and has an efficient parameter reestimation procedure using training data. For future research, we must explore the appropriate stochastic modeling techniques for the dialogue process, considering expressive power, computational cost and precision of parameter estimation, and the amount of training data available. Moreover, in order to increase the amount of training data available, we have to develop an automatic tagger for utterance types.

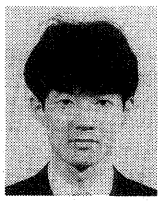
Acknowledgements The authors would like to thank Dr. Kurematsu, and all the members of ATR Interpreting Telephony Research Laboratories, for their constant help and fruitful discussions.

References

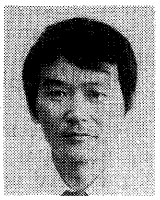
1) Austin, J.: *How to Do Things with Words*, Oxford University Press, New York (1962).
2) Ehara, T., Ogura, K. and Morimoto, T.: ATR Dialogue Database, ICSLP-90, pp. 1093-1096 (1990).
3) Grosz, B. J. and Sidner, C. L.: Attention, Intentions, and the Structure of Discourse, *Computational Linguistics*, Vol. 12, No. 3, pp. 175-204 (1986).
4) Hasegawa, T.: Rule Application Control Method in Lexicon-Driven Transfer Model of a Dialogue Translation System, *ECAI-90*, pp. 336-338 (1990).
5) Hauptmann, A. G., Young, S. R. and Ward, W.

H.: Using Dialog-Level Knowledge Sources to Improve Speech Recognition, *AAAI-88*, pp. 729-733 (1988).
6) Jelinek, F.: Self-Organized Language Modeling for Speech Recognition, IBM T. J. Watson Research Center, Unpublished (1985).
7) Kita, K., Takezawa, T., Hosaka, J., Ehara, T. and Morimoto, T.: Continuous Speech Recognition Using Two-level LR Parsing, *ICSLP-90*, pp. 905-908 (1990).
8) Kume, M., Sato, G. K. and Yoshimoto, K.: A Descriptive Framework for Translating Speaker's Meaning—Towards a Dialogue Translation System between Japanese and English, *EACL-89*, pp. 264-271 (1989).
9) Magerman, D. M. and Marcus, M. P.: Parsing a Natural Language Using Mutual Information Statistics, *AAAI-90*, pp. 984-989 (1990).
10) Morimoto, T., Suzuki, S., Takezawa, T., Kikui, G., Nagata, M. and Tomokiyo, M.: A Spoken Language Translation System: SL-TRANS2, *COLING-92*, pp. 1048-1052 (1992).
11) Nagata, M.: An Empirical Study on Rule Granularity and Unification Interleaving Toward an Efficient Unification-Based Parsing System, *COLING-92*, pp. 177-183 (1992).
12) Nagata, M.: Using Pragmatics to Rule Out Recognition Errors in Cooperative Task-Oriented Dialogues, *ICSLP-92*, pp. 647-650 (1992).
13) Searle, J.: *Speech Acts*, Cambridge University Press, Cambridge (1969).
14) Walker, M. and Whittaker, S.: Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation, *ACL-90*, pp. 70-78 (1990).
15) Yamaoka, T. and Iida, H.: Dialogue Interpretation Model and its Application to Next Utterance Prediction for Spoken Language Processing, *EUROSPEECH-91*, pp. 849-852 (1991).
16) Young, S. and Matessa, Y.: Using Pragmatic and Semantic Knowledge to Correct Parsing of Spoken Language Utterances, *EUROSPEECH-91*, pp. 223-227 (1991).

(Received April 5, 1993)
(Accepted January 13, 1994)



Masaaki Nagata received the B. E. and M. E. degrees in information Science from Kyoto University, Kyoto, Japan, in 1985 and 1987, respectively. In 1987, he joined NTT Communications and Information Processing Laboratories. From 1989 to 1993, he worked at ATR Interpreting Telephony Research Laboratories, Kyoto, Japan, and was engaged in the research and development of speech-to-speech translation systems. In 1993, he moved to NTT Network Information Systems Labotratories. His current research interests include speech recognition, natural language processing, and the integration of the two. He is a member of IPSJ, the Japanese Society for Artificial Intelligence, and Association for Computational Linguistics.



Tsuyoshi Morimoto received the B. E. and M. E. degrees in Electronics Engineering from Kyushu University, Fukuoka, Japan, in 1968 and 1970, respectively. In 1970, he joined the Electrical Communication Laboratories of NTT. He was engaged in the research and development of operating system and database retrieval. In 1987, he moved to ATR Interpreting Telephony Research Laboratories, Kyoto, Japan, and is currently with ATR Interpreting Telecommunications Research Laboratories. He is the Head of the Department-4. His research interests are the integration of speech recognition and language processing, and natural language understanding. He is a member of IPSJ, and the Japanese Society for Artificial Intelligence.