An Efficient Statistical Speech Act Type Tagging System for Speech Translation Systems

Hideki Tanaka and Akio Yokoo
ATR Interpreting Telecommunications Research Laboratories
2–2, Hikaridai, Seika–cho, Soraku–gun, Kyoto, 619–0288, Japan
{tanakah|ayokoo}@itl.atr.co.jp

Abstract

This paper describes a new efficient speech act type tagging system. This system covers the tasks of (1) segmenting a turn into the optimal number of speech act units (SA units), and (2) assigning a speech act type tag (SA tag) to each SA unit. Our method is based on a theoretically clear statistical model that integrates linguistic, acoustic and situational information. We report tagging experiments on Japanese and English dialogue corpora manually labeled with SA tags. We then discuss the performance difference between the two languages. We also report on some translation experiments on positive response expressions using SA tags.

1 Introduction

This paper describes a statistical speech act type tagging system that utilizes linguistic, acoustic and situational features. This work can be viewed as a study on automatic "Discourse Tagging" whose objective is to assign tags to discourse units in texts or dialogues. Discourse tagging is studied mainly from two different viewpoints, i.e., linguistic and engineering viewpoints. The work described here belongs to the latter group. More specifically, we are interested in automatically recognizing the speech act types of utterances and in applying them to speech translation systems.

Several studies on discourse tagging to date have been motivated by engineering applications. The early studies by Nagata and Morimoto (1994) and Reithinger and Maier (1995) showed the possibility of predicting dialogue act tags for next utterances with statistical methods. These studies, however, presupposed properly segmented utterances, which is not a realistic assumption. In contrast to this assumption, automatic utterance segmentation (or discourse segmentation) is desired here.

Discourse segmentation in linguistics, whether manual or automatic, has also received keen attention because such segmentation provides the foundation of higher discourse structures (Grosz and Sidner, 1986).

Discourse segmentation has also received keen attention from the engineering side because the natural language processing systems that follow the speech recognition system are designed to accept linguistically meaningful units (Stolcke and Shriberg, 1996). There has been a lot of research following this line such as (Stolcke and Shriberg, 1996) (Cettolo and Falavigna, 1998), to only mention a few.

We can take advantage of these studies as a preprocess for tagging. In this paper, however, we propose a statistical tagging system that optimally performs segmentation and tagging at the same time. Previous studies like (Litman and Passonneau, 1995) have pointed out that the use of a multiple information source can contribute to better segmentation and tagging, and so our statistical model integrates linguistic, acoustic and situational information.

The problem can be formalized as a search problem on a word graph, which can be efficiently handled by an extended dynamic programming algorithm. Actually, we can efficiently find the optimal solution without limiting the search space at all.

The results of our tagging experiments involving both Japanese and English corpora indicated a high performance for Japanese but a considerably lower performance for the English corpora. This work also reports on the use of speech act type tags for translating Japanese and English positive response expressions. Positive responses quite often appear in task–oriented dialogues like those in our tasks. They are often highly ambiguous and problematic in speech translation. We will show that these expressions can be effectively translated with the help of dialogue information, which we call speech act type tags.

2 The Problems

In this section, we briefly explain our speech act type tags and the tagged data and then formally define the tagging problem.

2.1 Data and Tags

The data used in this study is a collection of transcribed dialogues on a travel arrangement task between Japanese and English speakers mediated by interpreters (Morimoto et al., 1994). The transcriptions were separated by language, i.e., English and Japanese, and the resultant two corpora share the same content. Both transcriptions went through morphological analysis, which was manually checked. The transcriptions have clear turn boundaries (TB's).

Some of the Japanese and English dialogue files were manually segmented into speech act units (SA units) and assigned with speech act type tags (SA tags). The SA tags represent a speaker's intention in an utterance, and is more or less similar to the traditional illocutionary force type (Searle, 1969).

The SA tags for the Japanese language were based on the set proposed by Seligman et al. (1994) and had 29 types. The English SA tags were based on the Japanese tags, but we redesigned and reduced the size to 17 types. We believed that an excessively detailed tag classification would decrease the inter-coder reliability and so pruned some detailed tags.[1] The following lines show an example of the English tagged dialogues. Two turns uttered by a hotel clerk and a customer were segmented into SA units and assigned with SA tags.

<clerk's turn>
Hello, (*expressive*)
New York City Hotel, (*inform*)
may I help you ? (*offer*)
<customer(interpreter)'s turn>
Hello, (*expressive*)
my name is Hiroko Tanaka (*inform*)
and I would like to make a reservation for
a room at your hotel. (*desire*)

The tagging work to the dialogue was conducted by experts who studied the tagging manual beforehand. The manual described the tag definitions and turn segmentation strategies and gave examples. The work involved three experts for the Japanese corpus and two experts for the English corpus.[2]

The result was checked and corrected by one expert for each language. Therefore, since the work was done by one expert, the inter-coder tagging instability was suppressed to a minimum. As the result of the tagging, we obtained 95 common dialogue files with SA tags for Japanese and English and used them in our experiments.

[1] Japanese tags, for example, had four tags mainly used for dialogue endings: *thank, offer-follow-up, good-wishes, and farewell*, most of which were reduced to *expressive* in English.

[2] They did not listen to the recorded sounds in either case.

2.2 Problem Formulation

Our tagging system assumes an input of a word sequence for a dialogue produced by a speech recognition system. The word sequence is accompanied with clear turn boundaries. Here, the words do not contain any punctuation marks. The word sequence can be viewed as a sequence of quadruples:

$$\cdots(w_{i-1}, \boldsymbol{l}_{i-1}, \boldsymbol{a}_{i-1}, \boldsymbol{s}_{i-1}), (w_i, \boldsymbol{l}_i, \boldsymbol{a}_i, \boldsymbol{s}_i)\cdots$$

where w_i represents a surface wordform, and each vector represents the following additional information for w_i.

$\boldsymbol{l}_i$: canonical form and part of speech of w_i (linguistic feature)
$\boldsymbol{a}_i$: pause duration measured milliseconds after w_i (acoustic feature)
$\boldsymbol{s}_i$: speaker's identification for w_i such as clerk or customer (situational feature)

Therefore, an utterance like *Hello I am John Phillips and ...* uttered by a *customer* is viewed as a sequence like

(Hello, (hello, INTER), 100, customer), (I,(i, PRON),0, customer)), (am, (be, BE), 0, customer) ....

From here, we will denote a word sequence as $\boldsymbol{W} = w_1, w_2, \ldots w_i, \ldots, w_n$ for simplicity. However, note that $\boldsymbol{W}$ is a sequence of quadruples as described above.

The task of speech act type tagging in this paper covers two tasks: (1) segmentation of a word sequence into the optimal number of SA units, and (2) assignment of an SA tag to each SA unit. Here, the input is a word sequence with clear TB's, and our tagger takes each turn as a process unit. [3]

In this paper, an SA unit is denoted as u and the sequence is denoted as $\boldsymbol{U}$. An SA tag is denoted as t and the sequence is denoted as $\boldsymbol{T}$. x_s^e represents a sequence of x starting from s to e. Therefore, t_1^j represents a tag sequence from 1 to j.

The task is now formally addressed as follows: find the best SA unit sequence $\boldsymbol{U}$ and tag sequence $\boldsymbol{T}$ for each turn when a word sequence $\boldsymbol{W}$ with clear TB's is given. We will treat this problem with the statistical model described in the next section.

3 Statistical Model

The problem addressed in Section 2 can be formalized as a search problem in a word graph that holds all possible combinations of SA units in a turn. We take a probabilistic approach to this problem, which formalizes it as finding a path $(\hat{\boldsymbol{U}}, \hat{\boldsymbol{T}})$ in the word graph that maximizes the probability $P(\boldsymbol{U}, \boldsymbol{T} \mid \boldsymbol{W})$.

[3] Although we do not explicitly represent TB's in a word sequence in the following discussions, one might assume virtual TB markers like @ in the word sequence.

This is formally represented in equation (1). This probability is naturally decomposed into the product of two terms as in equation (3). The first probability in equation (3) represents an arbitrary word sequence constituting one SA unit u_j, given $\boldsymbol{h}_j$ (the history of SA units and tags from the beginning of a dialogue, $\boldsymbol{h}_j = u_1^{j-1}, t_1^{j-1}$) and input $\boldsymbol{W}$. The second probability represents the current SA unit u_j bearing a particular SA tag t_j, given u_j, $\boldsymbol{h}_j$, and $\boldsymbol{W}$.

$$(\hat{\boldsymbol{U}}, \hat{\boldsymbol{T}}) = \underset{\boldsymbol{U},\boldsymbol{T}}{\operatorname{argmax}} P(\boldsymbol{U}, \boldsymbol{T} \mid \boldsymbol{W}), \quad (1)$$

$$= \underset{\boldsymbol{U},\boldsymbol{T}}{\operatorname{argmax}} \prod_{j=1}^{k} P(u_j, t_j \mid \boldsymbol{h}_j, \boldsymbol{W}), \quad (2)$$

$$= \underset{\boldsymbol{U},\boldsymbol{T}}{\operatorname{argmax}} \prod_{j=1}^{k} P(u_j \mid \boldsymbol{h}_j, \boldsymbol{W}) \times P(t_j \mid u_j, \boldsymbol{h}_j, \boldsymbol{W}). \quad (3)$$

We call the first term "unit existence probability" P_E and the second term "tagging probability" P_T. Figure 1 shows a simplified image of the probability calculation in a word graph, where we have finished processing the word sequence of w_1^{s-1}.

Now, we estimate the probability for the word sequence w_s^{s+p-1} constituting an SA unit u_j and having a particular SA tag t_j. Because of the problem of sparse data, these probabilities are hard to directly estimate from the training corpus. We will use the following approximation techniques.

3.1 Unit Existence Probability

The probability of unit existence P_E is actually equivalent to the probability that the word sequence $w_s, \ldots, w_{s+p-1}$ exists as one SA unit given $\boldsymbol{h}_j$ and $\boldsymbol{W}$ (Fig. 1).

We then approximate P_E by

$$P_E \simeq P(B_{w_{s-1},w_s} = 1 \mid \boldsymbol{h}_j, \boldsymbol{W}) \times P(B_{w_{s+p-1},w_{s+p}} = 1 \mid \boldsymbol{h}_j, \boldsymbol{W}) \times \prod_{m=s}^{s+p-2} P(B_{w_m,w_{m+1}} = 0 \mid \boldsymbol{h}_j, \boldsymbol{W}), \quad (4)$$

where the random variable $B_{w_x,w_{x+1}}$ takes the binary values 1 and 0. A value of 1 corresponds to the existence of an SA unit boundary between w_x and w_{x+1}, and a value of 0 to the non-existence of an SA unit boundary. P_E is approximated by the product of two types of probabilities: for a word sequence break at both ends of an SA unit and for a non-break inside the unit. Notice that the probabilities of the former type adjust an unfairly high probability estimation for an SA unit that is made from a short word sequence.

The estimation of P_E is now reduced to that of $P(B_{w_x,w_{x+1}} \mid \boldsymbol{h}_j, \boldsymbol{W})$. This probability is estimated by a probabilistic decision tree and we have

$$P(B_{w_x,w_{x+1}} \mid \boldsymbol{h}_j, \boldsymbol{W}) \simeq P(B_{w_x,w_{x+1}} \mid \Phi_E(\boldsymbol{h}_j, \boldsymbol{W})),$$

where Φ_E is a decision tree that categorizes $\boldsymbol{h}_j, \boldsymbol{W}$ into equivalent classes (Jelinek, 1997). We modified C4.5 (Quinlan, 1993) style algorithm to produce probability and used it for this purpose. The decision tree is known to be effective for the data sparseness problem and can take different types of parameters such as discrete and continuous values, which is useful since our word sequence contains both types of features.

Through preliminary experiments, we found that $\boldsymbol{h}_j$ (the past history of tagging results) was not useful and discarded it. We also found that the probability was well estimated by the information available in a short range of r around w_x, which is stored in $\boldsymbol{W}$. Actually, the attributes used to develop the tree were in $\boldsymbol{W}' = w_{x-r+1}^{x+r}$: surface wordforms for w_{x-r+1}^{x+r}, parts of speech for w_{x-r+1}^{x+r}, and the pause duration between w_x and w_{x+1}. The word range r was set from 1 to 3 as we will report in sub-section 5.3.

As a result, we obtained the final form of P_E as

$$P_E \simeq P(B_{w_{s-1},w_s} = 1 \mid \Phi_E(\boldsymbol{W}')) \times P(B_{w_{s+p-1},w_{s+p}} = 1 \mid \Phi_E(\boldsymbol{W}')) \times \prod_{m=s}^{s+p-2} P(B_{w_m,w_{m+1}} = 0 \mid \Phi_E(\boldsymbol{W}')) \quad (5)$$

3.2 Tagging Probability

The tagging probability P_T was estimated by the following formula utilizing a decision tree Φ_T. Two functions named f and g were also utilized to extract information from the word sequence in u_j.

$$P_T \simeq P(t_j \mid \Phi_T(f(u_j), g(u_j), t_{j-1}, \ldots, t_{j-m})) \quad (6)$$

As this formula indicates, we only used information available with the u_j and m histories of SA tags in $\boldsymbol{h}_j$. The function $f(u_j)$ outputs the speaker's identification of u_j. The function $g(u_j)$ extracts cue words for the SA tags from u_j using a cue word list. The cue word list was extracted from a training corpus that was manually labeled with the SA tags. For each SA tag, the 10 most dependent words were extracted with a χ^2-test. After converting these into canonical forms, they were conjoined.

To develop a statistical decision tree, we used an input table whose attributes consisted of a cue word list, a speaker's identification, and m previous tags. The value for each cue word was a binary value, where 1 was set when the utterance u_j contained the word, or otherwise 0. The effect of $f(u_j)$, $g(u_j)$, and length m for the tagging performance will be reported in sub-section 5.3.

4 Search Method

A search in a word graph was conducted using the extended dynamic programming technique proposed

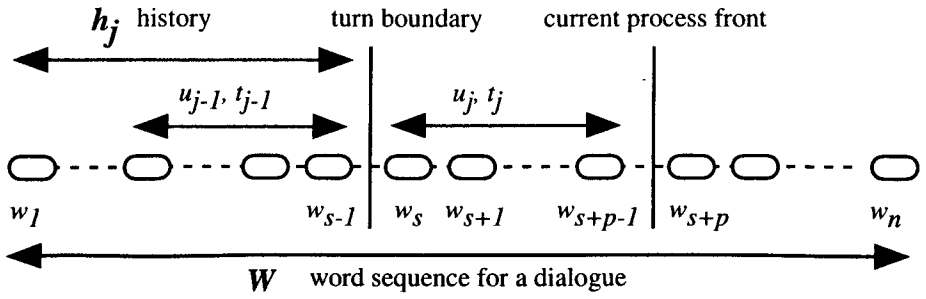

Figure 1: Probability calculation.

by Nagata (1994). This algorithm was originally developed for a statistical Japanese morphological analyzer whose tasks are to determine boundaries in an input character sequence having no separators and to give an appropriate part of speech tag to each word, i.e., a character sequence unit. This algorithm can handle arbitrary lengths of histories of pos tags and words and efficiently produce n–best results.

We can see a high similarity between our task and Japanese morphological analysis. Our task requires the segmentation of a word sequence instead of a character sequence and the assignment of an SA tag instead of a pos tag.

The main difference is that a word dictionary is available with a morphological analyzer. Thanks to its dictionary, a morphological analyzer can assume possible morpheme boundaries. [4] Our tagger, on the other hand, has to assume that any word sequence in a turn can constitute an SA unit in the search. This difference, however, does not require any essential change in the search algorithm.

5 Tagging Experiments

5.1 Data Profile

We have conducted several tagging experiments on both the Japanese and English corpora described in sub–section 2.1. Table 1 shows a summary of the 95 files used in the experiments. In the experiments described below, we used morpheme sequences for input instead of word sequences and showed the corresponding counts.

The average number of SA units per turn was 2.68 for Japanese and 2.31 for English. The average number of boundary candidates per turn was 18 for Japanese and 12.7 for English. The number of tag types, the average number of SA units, and the average number of SA boundary candidates indicated that the Japanese data were more difficult to process.

[4] Also, the probability for the existence of a word can be directly estimated from the corpus.

Table 1: Counts in both corpora.

Counts	Japanese	English
Turn	2,020	2,020
SA unit	5,416	4,675
Morpheme	38,418	27,639
POS types	30	33
SA tag type	29	17

5.2 Evaluation Methods

We used "labeled bracket matching" for evaluation (Nagata, 1994). The result of tagging can be viewed as a set of labeled brackets, where brackets correspond to turn segmentation and their labels correspond to SA tags. With this in mind, the evaluation was done in the following way. We counted the number of brackets in the correct answer, denoted as R (reference). We also counted the number of brackets in the tagger's output, denoted as S (system). Then the number of matching brackets was counted and denoted as M (match). Thus, we could define the precision rate with M/S and the recall rate with M/R.

The matching was judged in two ways. One was "segmentation match": the positions of both starting and ending brackets (boundaries) were equal. The other was "segmentation+tagging match": the tags of both brackets were equal in addition to the segmentation match.

The proposed evaluation simultaneously confirmed both the starting and ending positions of an SA unit and was more severe than methods that only evaluate one side of the boundary of an SA unit. Notice that the precision and recall for the segmentation+tagging match is bounded by those of the segmentation match.

5.3 Tagging Results

The total tagging performance is affected by the two probability terms P_E and P_T, both of which contain the parameters in Table 2. To find the best param-

Table 2: Parameters in probability terms.

P_E	P_T
w_{x-r+1}^{x+r} r: word range	$f(u_j)$: speaker of u_j $g(u_j)$: cue words in u_j $t_{j-1} \ldots t_{j-m}$: previous SA tags

Table 3: Average accuracy for segmentation match.

Parameter	Recall rate %	Precision rate %
A	89.50	91.99
B	91.89	92.92
C	92.00	92.57
D	92.20	92.58

Table 4: T–scores for segmentation accuracies.

	Recall				Precision		
	A	B	C		A	B	C
B	2.84	–	–	B	1.25	–	–
C	2.71	0.12	–	C	0.83	0.44	–
D	2.57	0.28	0.17	D	0.74	0.39	0.01

Table 5: Average accuracy for seg.+tag. match.

Parameter	Recall rate %	Precision rate %
E	72.25	72.70
F	74.91	75.35
G	74.83	75.29
H	74.50	74.96

eter set and see the effect of each parameter, we conducted the following two types of experiments.

I Change the parameters for P_E with fixed parameters for P_T
The effect of the parameters in P_E was measured by the segmentation match.

II Change the parameters for P_T with fixed parameters for P_E
The effect of the parameters in P_T was measured by the segmentation+tagging match.

Now, we report the details with the Japanese set.

5.3.1 Effects of P_E with Japanese Data

We fixed the parameters for P_T as $f(u_j)$, $g(u_j)$, t_{j-1}, i.e., a speaker's identification, cue words in the current SA unit, and the SA tag of the previous SA unit. The unit existence probability was estimated using the following parameters.

(A): Surface wordforms and pos's of w_x^{x+1}, i.e., word range $r = 1$

(B): Surface wordforms and pos's of w_{x-1}^{x+2}, i.e., word range $r = 2$

(C): (A) with a pause duration between w_x, w_{x+1}

(D): (B) with a pause duration between w_x, w_{x+1}

Under the above conditions, we conducted 10–fold cross–validation tests and measured the average recall and precision rates in the segmentation match, which are listed in Table 3.

We then conducted t–tests among these average scores. Table 4 shows the t–scores between different parameter conditions. In the following discussions, we will use the following t–scores: $t_{\alpha=0.025}(18) = 2.10$ and $t_{\alpha=0.05}(18) = 1.73$.

We can note the following features from Tables 3 and 4.

- recall rate
  (B), (C), and (D) showed statistically significant (two–sided significance level of 5%, i.e., $t > 2.10$) improvement from (A). (D) did not show significant improvement from either (B) nor (C).
- precision rate
  Although (B) and (C) did not improve from (A) with a high statistical significance, we can observe the tendency of improvement. (D) did not show a significant difference from (B) or (C).

We can, therefore, say that (B) and (C) showed equally significant improvement from (A): expansion of the word range r from 1 to 2 and using pause information with word range 1. The combination of word range 2 and pause (D), however, did not show any significant differences from (B) or (C). We believe that the combination resulted in data sparseness.

5.3.2 Effects of P_T with Japanese Data

For the Type II experiments, we set the parameters for P_E as condition (C): surface wordforms and pos's of w_x^{x+1} and a pause duration between w_x and w_{x+1}. Then, P_T was estimated using the following parameters.

(E): Cue words in utterance u_j, i.e., $g(u_j)$

(F): (E) with t_{j-1}

(G): (E) with t_{j-1} and t_{j-2}

(H): (E) with t_{j-1} and a speaker's identification $f(u_j)$

The recall and precision rates for the segmentation+tagging match were evaluated in the same way as in the previous experiments. The results are shown in Table 5. The t–scores among these parameter setting are shown in Table 6. We can observe the following features.

- recall rate
  (F) and (G) showed an improvement from (E) with a two–sided significance level of 10% ($t >$

Table 6: T-scores for seg.+tag. accuracies.

	Recall				Precision		
	E	F	G		E	F	G
F	1.87	–	–	F	1.97	–	–
G	1.78	0.05	–	G	1.90	0.04	–
H	1.50	0.26	0.21	H	1.60	0.28	0.24

1.73). However, (G) and (H) did not show significant improvements from (F).

- precision rate
  Same as recall rate.

Here, we can say that t_{j-1} together with the cue words (F) played the dominant role in the SA tag assignment, and the further addition of history t_{j-2} (G) or the speaker's identification $f(u_j)$ (H) did not result in significant improvements.

5.3.3 Summary of Japanese Tagging Experiments

As a concise summary, the best recall and precision rates for the segmentation match were obtained with conditions (B) and (C): approximately 92% and 93%, respectively. The best recall and precision rates for the segmentation+tagging match were 74.91% and 75.35 %, respectively (Table 5 (F)). We consider these figures quite satisfactory considering the severeness of our evaluation scheme.

5.3.4 English Tagging Experiment

We will briefly discuss the experiments with English data. The English corpus experiments were similar to the Japanese ones. For the SA unit segmentation, we changed the word range r from 1 to 3 while fixing the parameters for P_T to (H), where we obtained the best results with word range $r = 2$, i.e., (B). The recall rate was 71.92% and the precision rate was 78.10%. [5]

We conducted the exact same tagging experiments as the Japanese ones by fixing the parameter for P_E to (B). Experiments with condition (H) showed the best score: the recall rate was 53.17% and the precision rate was 57.75%. We obtained lower performance than that for Japanese. This was somewhat surprising since we thought English would be easier to process. The lower performance in segmentation affected the total tagging performance. We will further discuss the difference in section 7.

6 Application of SA tags to speech translation

In this section, we will briefly discuss an application of SA tags to a machine translation task. This is one of the motivations of the automatic tagging research described in the previous sections. We actually dealt with the translation problem of positive responses appearing in both Japanese and English dialogues.

Japanese positive responses like *Hai* and *Soudesuka*, and the English ones like *Yes* and *I see* appear quite often in our corpus. Since our dialogues were collected from the travel arrangement domain, which can basically be viewed as a sequence of a pair of questions and answers, they naturally contain many of these expressions.

These expressions are highly ambiguous in word-sense. For example, *Hai* can mean *Yes* (*accept*), *Uh huh* (*acknowledgment*), *hello* (*greeting*) and so on. Incorrect translation of the expression could confuse the dialogue participants. These expressions, however, are short and do not contain enough clues for proper translation in themselves, so some other contextual information is inevitably required.

We assume that SA tags can provide such necessary information since we can distinguish the translations by the SA tags in the parentheses in the above examples.

We conducted a series of experiments to verify if positive responses can be properly translated using SA tags with other situational information. We assumed that SA tags are properly given to these expressions and used the manually tagged corpus described in Table 1 for the experiments.

We collected Japanese positive responses from the SA units in the corpus. After assigning an English translation to each expression, we categorized these expressions into several representative forms. For example, the surface Japanese expression *Ee, Kekkou desu* was categorized under the representative form *Kekkou*.

We also made such data for English positive responses. The size of the Japanese and English data in representative forms (equivalent to SA unit) is shown in Table 7. Notice that 1,968 out of 5,416 Japanese SA units are positive responses and 1,037 out of 4,675 English SA units are positive responses. The Japanese data contained 16 types of English translations and the English data contained 12 types of Japanese translations in total.

We examined the effects of all possible combinations of the following four features on translation accuracy. We trained decision trees with the C4.5 (Quinlan, 1993) type algorithm while using these features (in all possible combinations) as attributes.

(I) Representative form of the positive response

(J) SA tag for the positive response

(K) SA tag for the SA unit previous to the positive response

(L) Speaker (Hotel/Clerk)

[5] Experiments with pause information were not conducted.

Table 7: Representation forms and the counts.

Japanese	freq.	English	freq.
Kekkou	69	I understand	6
Soudesu ka	192	Great	5
Hai	930	Okay	240
Soudesu	120	I see	136
Mochiron	7	All right	136
Soudesu ne	16	Very well	13
Shouchi	30	Certainly	27
Wakari-		Yes	359
mashita	304	Fine	52
Kashikomari-		Right	10
mashita	300	Sure	44
		Very good	9
Total	1,968	Total	1,037

Table 8: Accuracies with one feature.

Feature	J to E (%)	E to J (%)
I	54.83	46.96
J	51.73	34.33
K	73.02	55.35
L	40.09	37.80

We will show some of the results. Table 8 shows the accuracy when using one feature as the attribute. We can naturally assume that the use of feature (I) gives the baseline accuracy.

The result gives us a strange impression in that the SA tags for the previous SA units (K) were far more effective than the SA tags for the positive responses themselves (J). This phenomenon can be explained by the variety of tag types given to the utterances. A positive response expressions of the same representative form have at most a few SA tag types, say two, whereas the previous SA units can have many SA tag types. If a positive response expression possesses five translations, they cannot be translated with two SA tags.

Table 9 shows the best feature combinations at each number of features from 1 to 4. The best feature combinations were exactly the same for both translation directions, Japanese to English and vice versa. The percentages are the average accuracy obtained by the 10–fold cross–validation, and the t–score in each row indicates the effect of adding one feature from the upper row. We again admit a t–score that is greater than 2.01 as significant (two–sided significance level of 5 %).

The accuracy for Japanese translation was saturated with the two features (K) and (I). Further addition of any feature did not show any significant improvement. The SA tag for the positive responses did not work.

The accuracy for English translation was saturated with the three features (K), (I), and (L). The speaker's identification proved to be effective, unlike Japanese. This is due to the necessity of controlling politeness in Japanese translations according to the speaker. The SA tag for the positive responses did not work either.

Table 9: Best performance for each number of features.

Features	J to E (%)	t	E to J (%)	t
K	73.02	–	55.35	–
K,I	88.51	15.42	60.66	3.10
K,I,L	88.92	0.51	65.58	2.49
K,I,L,J	88.21	0.75	66.74	0.55

These results suggest that the SA tag information for the previous SA unit and the speaker's information should be kept in addition to representative forms when we implement the positive response translation system together with the SA tagging system.

7 Related Works and Discussions

We discuss the tagging work in this section. In subsection 5.3, we showed that Japanese segmentation into SA units was quite successful only with lexical information, but English segmentation was not that successful.

Although we do not know of any experiments directly comparable to ours, a recent work reported by Cettolo and Falavigna (1998) seems to be similar. In that paper, they worked on finding semantic boundaries in Italian dialogues with the "appointment scheduling task." Their semantic boundary nearly corresponds to our SA unit boundary. Cettolo and Falavigna (1998) reported recall and precision rates of 62.8% and 71.8%, respectively, which were obtained with insertion and deletion of boundary markers. These scores are clearly lower than our results with a Japanese segmentation match.

Although we should not jump to a generalization, we are tempted to say the Japanese dialogues are easier to segment than western languages. With this in mind, we would like to discuss our study.

First of all, was the manual segmentation quality the same for both corpora? As we explained in subsection 2.1, both corpora were tagged by experts, and the entire result was checked by one of them for each language. Therefore, we believe that there was not such a significant gap in quality that could explain the segmentation performance.

Secondly, which lexical information yielded such a performance gap? We investigated the effects of part–of–speech and morphemes in the segmentation

of both languages. We conducted the same 10-fold cross-validation tests as in sub-section 5.3 and obtained 82.29% (recall) and 86.16% (precision) for Japanese under condition (B'), which used only pos's in w_{x-1}^{x+2} for the P_E calculation. English, in contrast, marked rates of 65.63% (recall) and 73.35% (precision) under the same condition. These results indicated the outstanding effectiveness of Japanese pos's in segmentation. Actually, we could see some pos's such as "ending particle (shu-jyoshi)" which clearly indicate sentence endings and we considered that they played important roles in the segmentation. English, on the other hand, did not seem to have such strong segment indicating pos's. Although lexical information is important in English segmentation (Stolcke and Shriberg, 1996), what other information can help improve such segmentation?

Hirschberg and Nakatani (1996) showed that prosodic information helps human discourse segmentation. Litman and Passonneau (1995) addressed the usefulness of a "multiple knowledge source" in human and automatic discourse segmentation. Venditti and Swerts (1996) stated that the intonational features for many Indo-European languages help cue the structure of spoken discourse. Cettolo and Falavigna (1998) reported improvements in Italian semantic boundary detection with acoustic information. All of these works indicate that the use of acoustic or prosodic information is useful, so this is surely one of our future directions.

The use of higher syntactical information is also one of our directions. The SA unit should be a meaningful syntactic unit, although its degree of meaningfulness may be less than that in written texts. The goodness of this aspect can be easily incorporated in our probability term P_E.

8 Conclusions

We have described a new efficient statistical speech act type tagging system based on a statistical model used in Japanese morphological analyzers. This system integrates linguistic, acoustic, and situational features and efficiently performs optimal segmentation of a turn and tagging. From several tagging experiments, we showed that the system segmented turns and assigned speech act type tags at high accuracy rates when using Japanese data. Comparatively lower performance was obtained using English data, and we discussed the performance difference. We also examined the effect of parameters in the statistical models on tagging performance. We finally showed that the SA tags in this paper are useful in translating positive responses that often appear in task-oriented dialogues such as those in ours.

Acknowledgment

The authors would like to thank Mr. Yasuo Tanida for the excellent programming works and Dr. Seiichi Yamamoto for stimulus discussions.

References

M. Cettolo and D. Falavigna. 1998. Automatic detection of semantic boundaries based on acoustic and lexical knowledge. In *ICSLP '98*, volume 4, pages 1551–1554.

B. J. Grosz and C. L. Sidner. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3):175–204, July–September.

J. Hirschberg and C. H. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *34th Annual Meeting of the Association for the Computational Linguistics*, pages 286–293.

F. Jelinek, 1997. *Statistical Methods for Speech Recognition*, chapter 10. The MIT Press.

D. J. Litman and R. J. Passonneau. 1995. Combining multiple knowledge sourses for discourse segmentation. In *33rd Annual Meeting of the Association for the Computational Linguistics*, pages 108–115.

T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y. Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi, and Y. Yamazaki. 1994. A speech and language database for speech translation research. In *ICSLP '94*, pages 1791–1794.

M. Nagata and T. Morimoto. 1994. An information-theoretic model of discourse for next utterance type prediction. *Transactions of Information Processing Society of Japan*, 35(6):1050–1061.

M. Nagata. 1994. A stochastic Japanese morphological analyzer using a forward-DP and backward-A* N-best search algorithm. In *Proceedings of Coling94*, pages 201–207.

J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

N. Reithinger and E. Maier. 1995. Utilizing statistical dialogue act processing in verbmobil. In *33rd Annual Meeting of the Associations for Computational Linguistics*, pages 116–121.

J. R. Searle. 1969. *Speech Acts*. Cambridge University Press.

M. Seligman, L. Fais, and M. Tomokiyo. 1994. A bilingual set of communicative act labels for spontaneous dialogues. Technical Report TR-IT-0081, ATR-ITL.

A. Stolcke and E. Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *ICSLP '96*, volume 2, pages 1005–1008.

J. Venditti and M. Swerts. 1996. Intonational cues to discourse structure in Japanese. In *ICSLP '96*, volume 2, pages 725–728.