PROBABILISTIC DIALOGUE ACT EXTRACTION FOR CONCEPT BASED MULTILINGUAL TRANSLATION SYSTEMS

Toshiaki Fukada†‡ *Detlef Koll*† *Alex Waibel*† *Kouichi Tanigaki*‡

†Interactive Systems Laboratory, Carnegie Mellon University, Pittsburgh, USA

‡ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan

Email: fukada@itl.atr.co.jp koll@cs.cmu.edu ahw@cs.cmu.edu tanigaki@itl.atr.co.jp

ABSTRACT

This paper describes a probabilistic method for dialogue act (DA) extraction for concept-based multilingual translation systems. A DA is a unit of a semantic interlingua and it consists of speaker information, speech act, concept and argument. Probabilistic models for the extraction of speech acts or concepts are trained as speech act or concept dependent word n-gram models. The proposed method is evaluated on DA-annotated English and Japanese databases. The experimental results show that the proposed method gives a better performance compared to the conventioanl grammar-based approach. In addition, the proposed method is much more robust for erroneous inputs obtained as speech recognition outputs.

1. INTRODUCTION

In the C-STAR (Consortium for Speech Translation Advanced Research) project, several sites of spoken language groups, i.e., at CMU, ATR , UKA, ETRI, IRST [1], etc. are developing multilingual speech-to-speech translation systems [1][2][3]. To facilitate multilingual translation, a limited number of *dialogue acts* (DAs) called the *interchange format* (IF) are being used as the interlingual protocol. Figure 1 illustrates a component diagram of an IF based speech-to-speech translation system. In this figure, the *IF Extractor* performs the tasks of analyzing recognized text and mapping the information into an IF. The text is analyzed by a robust parser with semantic grammars. Although a grammar-based approach does not require an IF-tagged database, it has several drawbacks; (1) it requires time and expertise to construct, and (2) it is difficult to write an analysis grammar for erroneous inputs such as speech recognition results.

Recently, many statistical understanding approaches have been proposed and satisfying results have been obtained [4][5][6]. Although these approaches have mainly been developed in the ATIS (Air Travel Information System) domain, few works have applied a statistical approach in the speech translation domain [7].

The C-STAR project has just started collecting IF-tagged multilingual databases [8]. In this paper, a probabilistic approach is applied for DA extraction in speech translation tasks. Although similar approaches have been proposed [9][7][10], this paper differs in the following points:

- the proposed method copes with more complicated problems (e.g., both the speech act and concept are extracted),

[1]Carnegie Mellon University (USA), ATR Interpreting Telecommunications Research Laboratories (Japan), University Karlsruhe (Germany), Electronics and Telecommunications Research Institute (Korea), Istituto per la Ricerca Scientifica e Tecnologica (Italy)

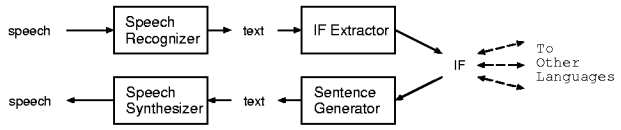

Figure 1. Components of a translation system.

- performances are compared to the conventional grammar-based approach,
- the proposed method is evaluated on two languages (i.e., English and Japanese).

In the following sections, we first explain the interchange format. In section **3.**, the probabilistic DA extraction scheme is described. Section **4.** shows performances for DA extraction both for the conventional grammar-based method and for the proposed method. Section **5.** gives a discussion of the presented work.

2. THE INTERCHANGE FORMAT

DAs indicate the intentions of speakers, and characterize the focus of the informational content of utterances. A scheme is developed for two-agent travel planning domain dialogues in which a travel agent and a customer are involved in various travel scenarios like hotel or flight reservation, ticket purchasing, transportation inquiry, tour or sightseeing information requesting, etc.

A DA consists of speaker information (agent or customer) and three representational levels indicating different aspects of the utterance: the *speech act*, the *concept*, and the *argument*. Table 1 shows examples of texts and their DAs. The speech acts capture the intentions of the speaker (i.e., whether the speaker performs the act of accepting, giving or requesting information, etc.). The concepts capture the informational focus of the utterance in question (i.e., whether the speaker is giving information about the availability of rooms, about a trip, a flight, etc.). The arguments denote the specific content of the utterance (e.g., whether the

Table 1. Examples of texts and their DAs.

text	`This is Rob.`
speech act	`introduce-self`
concept	*nil*
argument	`(person-name=rob)`
text	`The Pittsburgh arts festival is running from June seventh through the twenty third.`
speech act	`give-information`
concept	`temporal+event`
argument	`(event=pittsburgh_arts_festival, time=(start-time=(june, md7)), end-time=md23)`

speaker is giving information about single or double rooms, about one or two flights, etc.).

Currently, 26 speech acts (**s**), 64 concepts (**c**), and 77 arguments are defined. Speech act or concept may be adjoined to other speech acts or concepts in order to form new ones. However, there are constraints on the order, and so not all combinations are possible. Combinations of speech acts and concepts are also defined. The possible combinations of speech acts (**S**), concepts (**C**), and speech acts and concepts (**D**), are 68, 2090, and 24927, respectively. There are a number of utterances unable to be accounted for by the current inventory of DAs. These are either particularly complicated structures, false starts, or out-of-domain utterances. These utterances are annotated with "no-tag".

A DA is not assigned for each utterance but for each semantic phrase called a *semantic dialogue unit* (SDU). Therefore, some utterances are annotated with several DAs. In this paper, however, the segmentation problem (i.e., from an utterance to SDUs) is not considered. In the following section, we try to extract speech acts and concepts in a probabilistic way.

3. PROBABILISTIC DIALOGUE ACT EXTRACTION

3.1. Probabilistic framework

The goal of DA extraction is to to find the most likely IF, $\hat{\mathbf{I}}$, given a sequence of words $\mathbf{W}$, i.e., to maximize the probability $P(\mathbf{I}|\mathbf{W})$:

$$\hat{\mathbf{I}} = \underset{\mathbf{I}}{\operatorname{argmax}} P(\mathbf{I}|\mathbf{W}). \tag{1}$$

Using Bayes' Rule, the right-hand side of Eq.(1) can be written as

$$\begin{aligned}\hat{\mathbf{I}} &= \underset{\mathbf{I}}{\operatorname{argmax}} \frac{P(\mathbf{W}|\mathbf{I})P(\mathbf{I})}{P(\mathbf{W})} \\ &= \underset{\mathbf{I}}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{I})P(\mathbf{I}). \end{aligned} \tag{2}$$

The first term in Eq.(2), $P(\mathbf{W}|\mathbf{I})$, is the probability of a sequence of words, conditioned on the IF. The second term in Eq.(2), $P(\mathbf{I})$, is the *a priori* probability of generating $\mathbf{I}$.

3.2. DA model

3.2.1. Speech act models

First, we consider the extraction of the best combination of speech acts, $\hat{\mathbf{S}}$, among K ($K = 68$ in this paper) kinds of speech act combinations, $\mathbf{S}_k (1 \leq k \leq K)$, given $\mathbf{W}$. This can be found by simply replacing $\mathbf{I}$ with $\mathbf{S}$ in Eq.(2) as

$$\hat{\mathbf{S}} = \underset{\mathbf{S}_k}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{S}_k)P(\mathbf{S}_k). \tag{3}$$

Assume that we are given the sequence of words $\mathbf{W} = \{w_1, \ldots, w_T\}$ corresponding to one SDU. Suppose also that the conditional independence of w_t given the combination of speech acts, $\mathbf{S}_k = \{s_{k,1}, \ldots, s_{k,I_k}\}$, $P(\mathbf{W}|\mathbf{S}_k)$ can be computed as

$$P(\mathbf{W}|\mathbf{S}_k) = \prod_{t=1}^{T} P(w_t|\mathbf{S}_k) \tag{4}$$

$$= \prod_{t=1}^{T} \sum_{i=1}^{I_k} P(w_t|s_{k,i})P(s_{k,i}|\mathbf{S}_k), \tag{5}$$

where $s_{k,i}$ is an element of $\mathbf{s}$. Under the two assumptions, i.e., the conditional independence of $s_{k,i}$ given $\mathbf{S}_k$, and $P(w_t|s_{k,i})$, $P(w_t|s_{k,i})$ can be obtained by counting the number of words for each speech act. $P(s_{k,i}|\mathbf{S}_k)$ can be given as $P(s_{k,i}|\mathbf{S}_k) = 1/I_k$. $P(\mathbf{S}_k)$ is obtained by counting the number of $\mathbf{S}_k$.

3.2.2. Concept models

Next, we consider the extraction of the best combination of concepts, $\hat{\mathbf{C}}$, among L ($L = 2090$ in this paper) kinds of concept combinations, $\mathbf{C}_l (1 \leq l \leq L)$, given $\mathbf{W}$. This can be obtained by

$$\hat{\mathbf{C}} = \underset{\mathbf{C}_l}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{C}_l)P(\mathbf{C}_l). \tag{6}$$

Under the assumption of the conditional independence of w_t given the combination of concepts, $\mathbf{C}_l = \{c_{l,1}, \ldots, c_{l,J_l}\}$, $P(\mathbf{W}|\mathbf{C}_l)$ can be computed as:

$$P(\mathbf{W}|\mathbf{C}_l) = \prod_{t=1}^{T} \sum_{j=1}^{J_l} P(w_t|c_{l,j})P(c_{l,j}|\mathbf{C}_l), \tag{7}$$

where $c_{l,j}$ is an element of $\mathbf{c}$. $P(w_t|c_{l,j})$ is obtained by counting the number of words for each concept and $P(c_{l,j}|\mathbf{C}_l) = 1/J_l$. $P(\mathbf{C}_l)$ is obtained by counting the number of $\mathbf{C}_l$.

3.2.3. Speech act and concept models

Finally, we consider the extraction of the best combination of speech acts and concepts, $\hat{\mathbf{D}}$, among the M ($M = 24927$ in this paper) kinds of speech act and concept pairs, $\mathbf{D}_m (1 \leq m \leq M)$, given $\mathbf{W}$. Under the assumption that speech acts and concepts are independent, in addition to the same assumptions described in **3.2.1.** and **3.2.2.**, $\hat{\mathbf{D}}$ can be obtained by

$$\begin{aligned}\hat{\mathbf{D}} &= \underset{\mathbf{D}_m}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{D}_m)P(\mathbf{D}_m) \\ &= \underset{\mathbf{S}_k,\mathbf{C}_l}{\operatorname{argmax}} \{P(\mathbf{W}|\mathbf{S}_k)P(\mathbf{S}_k) \\ &\quad \cdot P(\mathbf{W}|\mathbf{C}_l)P(\mathbf{C}_l) \cdot \delta(\mathbf{S}_k, \mathbf{C}_l)\}, \end{aligned} \tag{8}$$

where

$$\delta(\mathbf{S}_k, \mathbf{C}_l) = \begin{cases} 1.0, & \text{if combination } \mathbf{S}_k \text{ and } \mathbf{C}_l \text{ is defined} \\ 0.0, & \text{otherwise.} \end{cases} \tag{9}$$

The probabilities in Eq.(8) can be computed by using the speech act and the concept models described in **3.2.1.** and **3.2.2.**. The possible combinations (i.e., M) of speech acts and concepts are predefined by experts.

4. EXPERIMENTS

4.1. Conditions

The DA-annotated English and Japanese databases on travel arrangement tasks shown in Table 2 were used for our evaluation. 64 English dialogues were used for training and 50 other English dialogues were used for test. As for the Japanese database, 84 dialogues were used for training and 42 other dialogues were used for test. Note that the 64 English dialogues were collected at CMU, and the others were collected at ATR independently. That is, for English the training and test sets were collected at two different sites in two independent data collections, while for Japanese the sets were collected in one data collection effort. Hence, the experiment for the English database was a much more difficult task than for the Japanese database. Actually, the unknown words for the English and Japanese databases in the test sets were 7.5% and 1.2%, respectively.

For the comparison, a grammar-based parsing approach was evaluated on the English database as our conventional method. Here, the English analysis grammar had 3,168 rules. The same units (i.e., SDUs) were given for the grammar-based method. Note that the grammar was developed through an end-to-end evaluation (i.e., grading of the translated results). "no-tag" data was excluded both from the training and test.

Table 2. Size of SDUs (Size of words is shown in brackets).

	English	Japanese
training	1,742 (10,719)	3,584 (31,515)
test	1,105 (8,145)	1,902 (16,182)

$P(w_t|s_{k,i})$ in Eq.(4) and $P(w_t|c_{l,j})$ in Eq.(7) were trained as speech act and concept dependent word unigram models, respectively. Here, as the speaker information (i.e., agent or customer) can be used, agent and customer dependent speech act models, $\bar{P}_a(w_t|s_i)$ and $\bar{P}_c(w_t|s_i)$, were generated as follows.

$$\bar{P}_a(w_t|s_i) = k \cdot P_a(w_t|s_i) + (1-k) \cdot P(w_t|s_i) \quad (10)$$

$$\bar{P}_c(w_t|c_i) = k \cdot P_c(w_t|s_i) + (1-k) \cdot P(w_t|s_i), \quad (11)$$

where $P_a(w_t|s_i)$ and $P_c(w_t|c_i)$ were trained from the agent and the customer data, respectively. k is an interpolation factor for training robust models. The concept models were also generated in the same way. k was set to 0.4.

In the following experiment on the English database, three kinds of preprocessings were applied for the sequence of words in advance; (1) elimination of function or interjection words such as "a" or "uh", (2) categorization of words (e.g., `Rob`, `Harris` → `*person-name*`, `Sunday`, `Monday` → `*day-of-week*`, `room`, `rooms` → `ROOM`, `is`, `are` → `BE`), and (3) composition of words (e.g., `how many` → `HOWMANY`). We confirmed that these preprocessings achieved about a 1 ∼ 2% better performance compared to the plain texts for the test set.

4.2. Comparison with grammar-based approach

4.2.1. Baseline performance

The estimation performance (%Correct) for the speech act (SA), concept, and both (i.e., SA and concept) are shown in Table 3. From this table, the proposed method performs better than the conventional grammar-based method.

Table 3. Estimation performance on an English database (SA/Concept/SA-Concept) (%).

	conventional	proposed
training	75.3 / 84.7 / 69.8	82.3 / 81.3 / 68.5
test	50.1 / 55.0 / 37.8	58.8 / 57.5 / 39.7

4.2.2. Robustness for erroneous texts

As the probabilistic approach does not use strong constraints between words compared to the grammar-based approach, we expect the proposed method to be robust for erroneous word sequences. To confirm this, 78 erroneous sentences, which were obtained from speech recognition results for the test set, were evaluated. The results are shown in Table 4. The degradation in the estimation performance for the proposed method is much less than that for the conventional method. This result indicates that the proposed probabilistic approach is more robust for erroneous texts than the conventional grammar-based approach. Note that the reason the performance for the given transcriptions was much worse than those in Table 3, is due to the choice of the test set.

Table 4. Degradation of estimation performance by erroneous input (SA/Concept/SA-Concept) (%)

	conventional	proposed
transcribed	48.7 / 41.0 / 29.5	53.9 / 34.6 / 20.5
recognized	30.8 / 23.1 / 14.1	50.0 / 29.5 / 18.0
degradation	36.8 / 43.7 / 52.2	7.2 / 14.7 / 12.2

4.2.3. Error analysis

We have confirmed that the proposed method gave a better performance especially for erroneous texts. The grammar-based approach, however, is still advantageous to the proposed method, since all IFs (i.e., arguments in addition to speech acts and concepts) can not be obtained by the proposed method. One realistic application is to construct a hybrid system by combining the probabilistic and grammar-based approaches. Then, probabilistic information can be used in order to identify the DA among the several hypotheses obtained by the grammar-based method. However, no significant difference in performance was observed in a baseline experiment (e.g., 37.8% vs. 39.7% for SA-Concept). This implies that the proposed method will not give useful information for the grammar-based system when the error tendencies are similar. The confusion matrix for the SA-Concept of the test set is listed in Table 5. We can see from this table that the proposed method offers the possibility of providing useful information to the grammar-based system. This table indicates that the performance of the conventional method will be improved more than 10% (i.e., 115/1105) when these two methods are ideally combined.

Table 5. Confusion matrix (SA-Concept).

conventional	proposed	
	#correct	#error
#correct	324	94
#error	115	572

4.3. Evaluation on the Japanese database

The estimation performance on the Japanese database is shown in Table 6. The N-best performance for the test set is shown in Figure 7. 95% of the speech acts and 85% of the concepts were correctly estimated in the three best

Table 6. Estimation performance on a Japanese database (SA/Concept/SA-Concept) (%).

training	89.2 / 77.7 / 66.8
test	79.9 / 71.6 / 57.6

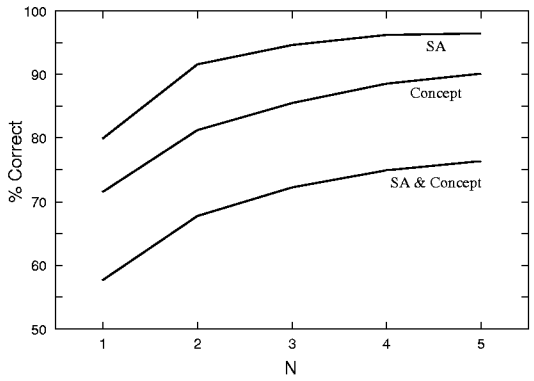

Table 7. N-best performance on the Japanese database.

hypotheses.

5. DISCUSSION

In section **3.**, we derived mathematical formulae under several independence assumptions. We discuss two techniques for relaxing these independence assumptions, i.e., independence of w_t, and $s_{k,i}$ and $c_{l,j}$.

5.1. Higher order N-gram model

The independence assumption of w_t is easily relaxed by using higher order N-gram models. That is, speech act or concept dependent word bigram models, $P(w_t|w_{t-1}, s_{k,i})$ or $P(w_t|w_{t-1}, c_{l,j})$, can be used instead of unigram models, $P(w_t|s_{k,i})$ or $P(w_t|c_{l,j})$. The difference in the estimation performance was investigated on the English and the Japanese databases. In this experiment, plain texts were used (i.e., no text filter was applied) and speaker information was not employed. Back-off smoothing was applied in the bigram training. The results are shown in Table 8. We can see from these results that the bigram models give a better performance for the speech acts, while no significant improvement is observed for the concepts. We consider that this is mainly because the word order is more important for the estimation of speech acts than concepts. In addition, the training data is insufficient for the 64 kinds of concept models compared to the 26 kinds of speech act models. Actually, for the Japanese case, a slight improvement was observed for the bigram model whose training data was more than two times larger than the English database. Table 9 shows the estimation performance for the amount of training data on the Japanese database. As the improvement of the bigram model is higher than that of the unigram model, significant differences can be expected when larger amounts of training data become available.

Table 8. Unigram vs. Bigram (SA/Concept/SA-Concept) (%).

	English	Japanese
unigram	51.7 / 54.3 / 35.2	79.6 / 70.1 / 56.2
bigram	56.9 / 51.1 / 32.9	81.2 / 70.5 / 57.9

Table 9. Estimation performance for the amount of training data (SA/Concept) (%).

#SDU	1000	2000	3584
unigram	73.3 / 62.1	76.7 / 67.1	79.6 / 70.1
bigram	73.9 / 60.3	77.5 / 65.6	81.2 / 70.5

5.2. Retraining as a mixture model

The assumption that each word is related to all relevant speech acts or concepts for a DA is not obviously appropriate. For example, in the last example in Table 1, "`The Pittsburgh arts festival`" is not related to the concept "`temporal`" and "`from June seventh through the twenty third`" is not related to the concept "`event`". Therefore, under this assumption, the probabilistic overlap among the speech act models or concept models becomes broad, since the speech acts or concepts that occur in the same DA share the words in that DA. This assumption can be relaxed by considering these models as an HMM (Hidden Markov Model) and training the probabilities as a mixture model with the EM (Expectation and Maximization) algorithm [11]. The EM algorithm attempts to maximize the expected likelihood $\mathbf{P} = \prod_{n=1}^{N} P(\mathbf{W}_n|\mathbf{I}_n)$, where N is the total number of SDUs in the training data. We expect that this leads to an improvement in the discrimination between speech acts or concepts.

As a preliminary experiment, the concept models were trained as a mixture HMM on the Japanese database. Note that $P(w_t|c_{l,j})$ and $P(c_{l,j}|\mathbf{C}_l)$ in Eq.(7) can be considered as an output probability and a transition probability in a mixture HMM, and these probabilities are trained by the EM algorithm. Here, we used a special concept of "general words" which was designed for taking general or irrelevant words for defined concepts such as "the" or "uh". In the training, an interjection was forced to be in the special concept. As a result, we observed a significant improvement from 70.1% to 76.2 %.

6. CONCLUSION

We have proposed a DA extraction method based on a probabilistic approach. The experimental results showed that the proposed method gives a better performance and is more robust for erroneous texts compared to the conventional grammar-based approach. The proposed method will also be applicable for the purpose of automatic (or semi-automatic) DA annotation which requires expertise and time.

As the current IF databases are quite small, language model adaptation will be a useful technique for improving the performance. The performance will additionally be improved by incorporating historical information [9][10][7] (e.g., $P(\mathbf{I}_n|\mathbf{I}_{n-1})$ instead of $P(\mathbf{I}_n)$), since DA assignment is not strictly SDU or utterance based: both the immediate and distant contexts are taken into account.

ACKNOWLEDGMENT

The authors wish to thank Dr. Seiichi Yamamoto, President of ATR Interpreting Telecommunications Research Laboratories, for giving us the opportunity to carry out this study. The English analysis grammar was developed by Christie Watson and Kavita Thomas. We would also like to thank all members in our speech groups at CMU and ATR for their helpful discussions.

REFERENCES

[1] B. Reaves, A. Nishino and T. Takezawa: "ATR-MATRIX: Implementation of a speech translation system," *Proc. Acoust. Soc. Japan Spring Meeting*, pp. 53–54, Mar. 1998.

[2] A. Lavie, L. Levin, P. Zhan, M. Taboada, D. Gates, M. Lapata, C. Clark, M. Broadhead and A. Waibel: "Expanding the domain of a multi-lingual speech-to-speech translation system," *Proc. Workshop on Spoken Language Translation*, ACL/EACL-97, July 1997.

[3] B. Angelini, M. Cettolo, A. Corazza, D. Falavigna and G. Lazzari: "Multilingual person to person communication at IRST," *Proc. ICASSP-97*, pp. 91–94, 1997.

[4] R. Kuhn and R. De Mori: "The application of semantic classification trees to natural language understanding," *IEEE Trans. on PAMI*, vol.17, no.5, pp. 449–460, May 1995.

[5] R. Schwartz, S. Miller, D. Stallard and J. Makhoul: "Hidden understanding models for statistical sentence understanding," *Proc. ICASSP-97*, pp. 1479–1482, 1997.

[6] K. Papineni, S. Roukos and R. Ward: "Maximum likelihood and discriminative training of direct translation models," *Proc. ICASSP-98*, pp. 189–192, 1998.

[7] N. Reithinger and M. Klesen: "Dialogue act classification using language models," *Proc. Eurospeech-97*, pp. 2235–2238, 1997.

[8] L. Levin, D. Gates, A. Lavie and A. Waibel: "An interlingua based on domain actions for machine translation of task-oriented dialogues," *Proc. ICSLP-98*, 1998.

[9] M. Woszczyna and A. Waibel: "Inferring linguistic structure in spoken language," *Proc. ICSLP-94*, pp. 847–850, 1994.

[10] Y.-Y. Wang and A. Waibel: "Statistical analysis of dialogue structure," *Proc. Eurospeech-97*, pp. 2703–2706, 1997.

[11] T. Imai, R. Schwartz, F. Kubala and L. Nguyen: "Improved topic discrimination of Broadcast News using a model of multiple simultaneous topics," *Proc. ICASSP-97*, pp. 727–730, 1997.