interspeech 2008 incorporating SST 08

Probabilistic Answer Selection Based on Conditional Random Fields for Spoken Dialog System

Yoshitaka Yoshimi, Ryota Kakitsuba, Yoshihiko Nankaku, Akinobu Lee, Keiichi Tokuda

Department of Computer Science and Engineering
Nagoya Institute of Technology, Japan

Abstract

A probabilistic answer selection for a spoken dialog system based on Conditional Random Fields (CRFs) is described. The probabilities of answers for a question is trained by CRFs based on the lexical and morphological properties of each word, the most likely answer against the recognized word sequence of question utterance will be chosen as the system output. Various set of feature functions were evaluated on the real data of a speech oriented information kiosk system, and it is shown that the morphological properties introduces positive effects on the response accuracy. Training with recognizer output of training database instead of manual transcription was also investigated. It was also shown that this proposed scheme can achieve higher accuracy than a conventional keyword-based answer selection.
Index Terms: spoken dialog system, probabilistic answer selection, CRFs

10.21437/Interspeech.2008-67

1. Introduction

In recent years, many challenges to build a spoken dialog system that truly appeals to everyone has been tackled. We have also been trying to develop a public information kiosk based on a spoken dialog system with an animated agents, that can guide visitors about public facilities, local geographies, tourist information, transportations, weather forecasts, and so on. This empirical study has shown us that, on such system, people tend to ask a simple short question instead of long sentence or conversation interaction, and people expect a broad range of information to be extracted. Therefore, such system should have an ability to answer each user's question one by one precisely.

To achieve such systems, it should be adequate to take a strategy of simple dialog management rather than an elaborated, tailored system. Spoken dialog system such as Call For Fire (CFF) [1], web-based multi-modal dialog interaction [2], or interactive information guidance system [3] are fit to perform on complicated task using dialog history. However, such dialog management may depends on the kind of task. This paper aims at a simple answering which does not use a dialog history.

A simple method of answer selection for spoken dialog system can be a keyword extraction from recognition result[4]. However, the set of keyword types and answer scoring method are manually tailored for the specific task empirically. An another system uses detection of pertinent information non-contextually and search pertinent information chunks[5]. Such system is typically rule-based, and many detailed knowledge should be prepared for a task.

This paper addresses a probabilistic answer selection for a public information kiosk system. Generally, it is hard to collect a certain amount of training data for a dialog system enough to train its probabilistic feature, and it causes shortage of training data. We adopt Conditional Random Fields (CRFs)[6] for the question-to-answer mapping as our first trial, since CRFs has been popular for discriminative models.

Generally, applying appropriate feature function sets for CRFs are essential problem. In this paper, we compare several feature function sets for input sentence as combinations of lexical and morphological information. We also evaluated using recognition result instead of manual transcription in training CRFs. The performance will also be compared with a conventional method based on keyword matching[4]. Effect of hyper parameter values of CRFs are also exploited.

2. Answer selection based on CRFs

This section describes a basic framework of the probabilistic answer selection for spoken dialog system. Given a task domain of a system, a conditional probability $P(a|\boldsymbol{W})$ represents the output probability of an answer a given a question $\boldsymbol{W}$, and the most likely answer $\hat{a}$ should be determined. Where, $\hat{a}$ can be written as follows:

$$\hat{a} = \underset{a \in \mathcal{A}}{\operatorname{argmax}} P(a|\boldsymbol{W}). \tag{1}$$

$\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_t]^\top$ is a question utterance text consisting of words and morpheme information obtained by the recognizer, for example "How's the weather tomorrow?" $\boldsymbol{w}_i = [w_{i,1}, w_{i,2}, \ldots, w_{i,l}]$ is lexical and morpheme information at $\boldsymbol{w}_i$. a is an answer index to the pre-defined answers, such as "It's fine today.". $\mathcal{A}$ is the candidate answer set as list of answering text for the task domain. Here, we assume a simple QA dialog disregarding its discourse.

We use CRFs for the probabilistic answer selection. CRF is a probabilistic framework widely used for labeling and segmenting sequential data. Also, a CRF performs well in natural language processing[7]. CRFs can be learned to maximize a conditional probability $P(a|\boldsymbol{W})$ to be written as:

$$P(a|\boldsymbol{W}) = \frac{1}{Z(\boldsymbol{W})} \exp\left(\boldsymbol{\Lambda} \cdot \boldsymbol{F}\left(\boldsymbol{W}, a\right)\right), \tag{2}$$

Where, $Z(\boldsymbol{W})$ is a normalization factor over all candidate paths, $\boldsymbol{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ is a learned weight for each feature function to be estimated from training data, $\boldsymbol{F} = [F_1(\boldsymbol{W}, a), F_2(\boldsymbol{W}, a), \ldots, F_K(\boldsymbol{W}, a)]$ is feature function vector, K is a number of feature function, and $F_k(\boldsymbol{W}, a)$ is feature function of the observations and the labels.

Estimation of the weight is typically performed by likelihood maximization. The weight follows log likelihood because we model the conditional distribution. If training data is $T = \{(\boldsymbol{W}^{(1)}, a^{(1)}), (\boldsymbol{W}^{(2)}, a^{(2)}), \ldots, (\boldsymbol{W}^{(N)}, a^{(N)})\}$, log

Accepted after peer review of full paper

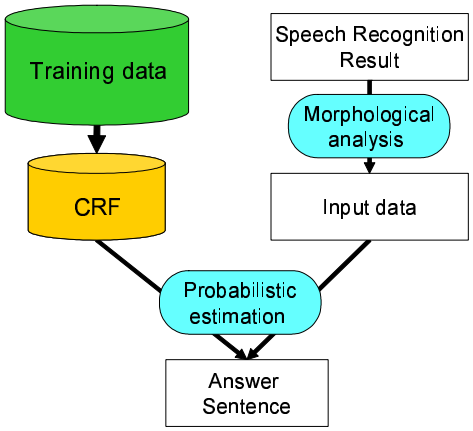

Figure 1: *System overview.*

likelihood is:

$$\mathcal{L}_{\mathbf{\Lambda}} = \sum_{j=1}^{N} \log P(a^{(j)}|\boldsymbol{W}^{(j)}). \quad (3)$$

Also, regularization of learned weight should be performed to avoid over-fitting. In this paper, Gaussian prior[8] is taken as a prior distribution as follows:

$$\mathcal{L}_{\mathbf{\Lambda}} = \sum_{j=1}^{N} \log P(a^{(j)}|\boldsymbol{W}^{(j)}) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2C}, \quad (4)$$

where, the parameter $C(> 0)$ is a hyper parameter of CRFs.

3. System

3.1. Overview

Figure 1 illustrates the answer selection system based on the CRF. First, the relation between question and answer is modeled using the CRF from training data. When the system operates, the system estimates the probability of most likely candidate answer when it receives an input text as a speech recognition result.

The training data is manual transcription of actual utterances spoken toward the system on real operation on public space. One can also use their recognition result as training data. In this case, the CRF may learn its probability considering the tendency of recognition errors. Also, the system can utilize N-best recognition result as test data.

3.2. Training data

Training data consists of transcribed questions and its answer index, in which the questions are annotated with their morpheme information. The answers are hand-labeled. An example of the training data is shown in Figure 2. Where, symbols *r, m, p0, p{1, 2, 3}, cf,* and *ct* are for answer, word class (POS), detailed word class{1, 2, 3}, conjugation form, and conjugation type, respectively. Furthermore, symbols *p{0, 1, 2, 3}* are hierarchical where *p{0}* is the top level. Also, symbols *c* and *x* are blank symbols which tells that there are no conjugation form and conjugation type for the word.

The training data for the CRF is annotated by morphological analysis of question texts. Also, an answer index for each training question is chosen manually from pre-defined answer set.

3.3. Feature functions

The feature function used in this study is word and morpheme information which are word class, detailed word class, conjugation form, and conjugation type. Detailed word class can classify words into narrow range if some words are the same word class. Also, a template of feature function packs type of features.

Table 1 shows the templates of feature functions and those groups. Where, symbols in the table corresponds to the top symbols of Figure 2. Feature functions are created from the templates and training data. For example, a kind of feature function $\langle$answer $\times$ word $\times$ word class$\rangle$ is made from a template $\langle r, m, p0\rangle$. Also, group names such as $R, M, G1, G2$ are bring together to some templates. For example, group $G1$ includes three kinds of templates, $\langle r, m, p0\rangle$, $\langle r, m, p0, p1\rangle$, and $\langle r, m, p0, p1, p2\rangle$. Several combinations of template groups are investigated to see the effect of each morpheme feature.

3.4. Answer selection algorithm

The algorithm of answer selection when the system receives speech recognition result is as follows.

1. The system get a morpheme sequence from speech recognition sentence using morphological analysis.
2. For each morpheme sequence, the system estimates probability for each candidate answers based on the CRF.
3. For each morpheme sequence, the system selects the most likely candidate answer.

when the system adopts N-best recognition result as input, the algorithm will continue as follows:

4. The system add up probability to candidate answer which have the highest probability of all as the probabilistic score over each recognition result.
5. The system determine candidate answer which have the highest score of all as the definite answer.

4. Experiments

4.1. Conditions

Q&A data collected by speech-oriented guidance system "Takemaru-kun"[4] was used to run experiment of the proposed probabilistic answer selection. This system is located at the front desk of a public city center, and aims to answer user's questions about its facilities, services, town information, traffic guidance and so on. This system has been installed at the North Community Center in Ikoma City, Nara Prefecture, Japan for more than four years. In this experiment, adult's utterances were used as test data. Table 2 shows the contents of the Q&A data sets.

Training data consists of utterances collected from November 2002 to October 2004 (excluding August 2003), in which only utterances that appeared more than twice are used. Test data are utterances on August 2003. The answer set consists of 180 answer sentences, and the system's response accuracy, i.e. the response correct rate, is calculated based on a one-to-one correspondence of user inputs to system responses.

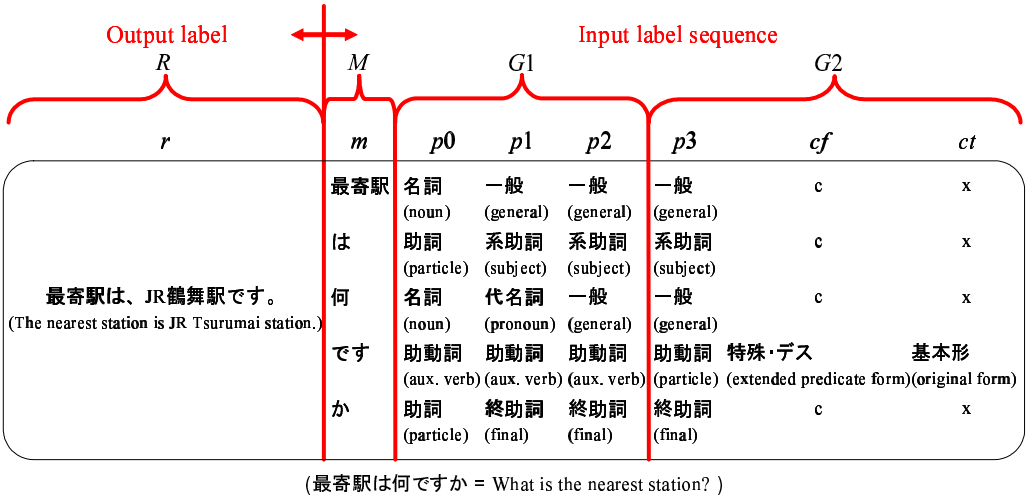

Figure 2: *An example of training data. Where, symbols r, m, p0, p{1, 2, 3}, cf, and ct mean answer, word class (POS), detailed word class{1, 2, 3}, conjugation form, and conjugation type. Furthermore, symbols p{0, 1, 2, 3} are hierarchical where p{0} is the top level.*

Table 1: *Feature templates.*

Group	Template
R	$\langle r \rangle$
M	$\langle r, m \rangle$
$G1$	$\langle r, m, p0 \rangle$
	$\langle r, m, p0, p1 \rangle$
	$\langle r, m, p0, p1, p2 \rangle$
$G2$	$\langle r, m, p0, p1, p2, p3 \rangle$
	$\langle r, m, p0, p1, p2, p3, cf \rangle$
	$\langle r, m, p0, p1, p2, p3, cf, ct \rangle$

Table 2: *Data sets.*

	Training data	Test data
Period (year.month)	2002.11∼2004.10 (Without 2003.08)	2003.08
# of utter.	13,487	1,091
Word acc.	91.1%	80.3%
N-best result	1-best	1/10-best

Table 3: *Comparison of templates.*

(Training data = Transcription)

	Template set	Response acc.[%]	
		Trans.	Recog.
(1)	R, M	76.6	71.7
(2)	R, M, M'	76.9	72.4
(3)	$R, G1$	77.0	72.7
(4)	$R, G1, G1'$	76.9	72.2
(5)	$R, G1, G2$	77.3	72.5
(6)	$R, G1, G2, G1', G2'$	77.1	72.1

4.2. Comparison of feature functions

To evaluate the appropriate feature functions, comparison of feature functions were performed. We used six kinds of feature template types from group names in Table 1 to train the CRF. For example, template set "$R, G1$", "$R, G1, G2$", and so on.

The result is shown in Table 3. Where, $\cdot'$ is the feature group for adjacent morpheme. Training data is manual transcription, and both transcription and 10-best recognition results are compared.

As a result, as shown as (1) and (5) in the table, the response rate was improved by 0.7% with morpheme information, regardless of the input transcription. It is shown that morpheme information should have positive effect as feature functions for answer selection. However, as shown at (3) with (5), 0.2% of deterioration was found by using all morpheme information when input data is recognition result. It means that, more detailed information should not be necessary, since using all morpheme information may sparse training.

Next, let us focus on the effect of incorporating word and morpheme context on the CRFs. In this comparative experimental result, as shown at (1) and (2), the word context information (feature group M) is vital as feature function. On the other hand, when the bars from (3) to (6) are compared, it is evident that morpheme-level context information (feature group $G1, G2$) did not contribute to the accuracy.

4.3. Comparison of training data

Next, the comparison of training data was performed. Table 4 shows the result. In this experiment, the feature set type of (5) on Table 3 was used as the feature template set. Both 1-best and 10-best recognition results are tested as input to the system.

The result shows that RA gains response accuracy when the training data is the automatic transcription which contains recognition errors. It suspects that we could achieve better performance when considering recognition errors by using recognition result as training data.

Table 4: *Comparison of training data.*

(Template set = Table 3 (5))

Training data	Response acc.[%]		
	Test data		
	Trans.	1-best	10-best
Transcription	77.3	72.0	72.5
Recognition result	75.3	72.7	73.5

Table 5: *Comparison with conventional method.*

Training data	Response acc.[%]	
	Test data	
	Trans.	Recog.
Conventional method	75.8	71.5
Proposed (Training data = Trans.)	77.3	72.5
Proposed (Training data = Recog.)	75.3	73.5

4.4. Comparison with conventional method

Finally, comparison with a conventional keyword-based method[4] were performed. Table 5 shows the result comparing the proposed method with feature template set (5) on Table 3 and 10-best input as the conventional system. The proposed method outperforms the conventional method by 1.5%. Also, RA indicates 1% of improvement when using proposed method, transcription as training data, and recognition result as test data. Furthermore, RA indicates 2% of improvement when using proposed method and recognition result as training and test data.

4.5. Comparison of hyper parameter values

Additionally, the appropriate value of hyper parameter C was exploited. Table 6 and Table 7 show the experimental result. The conditions are the same as previous section. As a result, RA reached its peak at $C = 1$ when using transcription as training data. However, as shown in Table 7, the system has a peak at different value of C when using recognition result as training data. Thus, appropriate value of C may depend on the input data.

5. Conclusion

A probabilistic answer selection for a spoken dialog system based on CRFs is described. The probabilities of answers toward a question is modeled by CRFs using transcriptions with lexical and morphological properties of each word. Experimental results showed that morpheme level information has positive effect to be included in the feature template set, but context information of the morphoneme resulted in less improvement than word-level context information. It is also shown that training CRFs with recognizer output that contains recognition errors can improve the response accuracy.

Furthermore, it was also shown that this proposed scheme can achieve higher accuracy than a conventional keyword-based scoring.

Future work should be dedicated to more robust training, integration with posterior probabilities from recognizer, or usage of N-best recognition result as training, and more broadened test on real system. We should also see its language-dependency.

Table 6: *Comparison of hyper parameter values C (Training data = Transcription).*

(Template set = Table 3 (5))

C	Response acc.[%]	
	Trans.	Recog.
0.1	71.9	68.3
1	77.3	72.5
10	76.4	72.4
20	76.5	72.4
40	76.4	72.1

Table 7: *Comparison of hyper parameter values C (Training data = Recognition result).*

(Template set = Table 3 (5))

C	Response acc.[%]	
	Trans.	Recog.
0.1	71.2	69.3
1	75.3	73.5
10	75.7	73.4
20	75.6	72.9
40	75.3	72.5

6. Acknowledgements

The authors would like to thank the Speech and Acoustics Processing Laboratory at NAIST, Japan to provide us the database.

7. References

[1] A. Roque, A. Leuski, V. Rangarajan, S. Robinson, A. Vaswani, S. Narayanan, and D. Traum, "Radiobot-CFF: A Spoken Dialog System for Military Training," *Proc. of ICSLP*, pp.477-480, 2006.

[2] A. Gruenstein, S. Seneff, and C. Wang, "Scalable and Portable Web-Based Multimodal Dialog Interaction with Geographical Databases," in *Proc. of ICSLP*, pp.453-456, 2006.

[3] T. Misu and T. Kawahara, "Speech-based Interactive Information Guidance System Using Question-Answering Technique," in *Proc. of ICASSP*, pp.145-148, 2007.

[4] R. Nisimura, Y. Nishihara, R. Tsurumi, A. Lee, H. Saruwatari, and K. Shikano, "Takemaru-kun: Speech-Oriented Information System for Real World Research Platform," in *International Workshop on Language Understanding and Agents for Real World Interaction*, pp.70-78, 2003.

[5] S. Rosset, O. Galibert, G. Illouz, and A. Max, "Integrating Spoken Dialog and Question Answering: the Ritel Project," in *Proc. of ICSLP*, pp.1914-1917, 2006.

[6] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of ICML*, pp.282-289, 2001.

[7] T.Kudo, K.Yamamoto, and Y.Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis," in *Proc. of EMNLP*, 2004.

[8] S.F. Chen and R. Rosenfeld, "A gaussian prior for smoothing maximum entropy models," Technical report, Carnegie Mellon University, 1999.