Speakers' Intention Prediction Using Statistics of Multi-level Features in a Schedule Management Domain

Donghyun Kim
Diquest Research Center
Diquest Inc.
Seoul, Korea
kdh2007@sogang.ac.kr

Hyunjung Lee
Computer Science & Engineering
Sogang University
Seoul, Korea
juvenile@sogang.ac.kr

Choong-Nyoung Seon
Computer Science & Engineering
Sogang University
Seoul, Korea
wilowisp@gmail.com

Harksoo Kim
Computer & Communications Engineering
Kangwon National University
Chuncheon, Korea
nlpdrkim@kangwon.ac.kr

Jungyun Seo
Computer Science & Engineering
Sogang University
Seoul, Korea
seojy@sogang.ac.kr

Abstract

Speaker's intention prediction modules can be widely used as a pre-processor for reducing the search space of an automatic speech recognizer. They also can be used as a pre-processor for generating a proper sentence in a dialogue system. We propose a statistical model to predict speakers' intentions by using multi-level features. Using the multi-level features (morpheme-level features, discourse-level features, and domain knowledge-level features), the proposed model predicts speakers' intentions that may be implicated in next utterances. In the experiments, the proposed model showed better performances (about 29% higher accuracies) than the previous model. Based on the experiments, we found that the proposed multi-level features are very effective in speaker's intention prediction.

1 Introduction

A dialogue system is a program in which a user and system communicate in natural language. To understand user's utterance, the dialogue system should identify his/her intention. To respond his/her question, the dialogue system should generate the counterpart of his/her intention by referring to dialogue history and domain knowledge. Most previous researches on speakers' intentions have been focused on intention identification techniques. On the contrary, intention prediction techniques have been not studied enough although there are many practical needs, as shown in Figure 1.

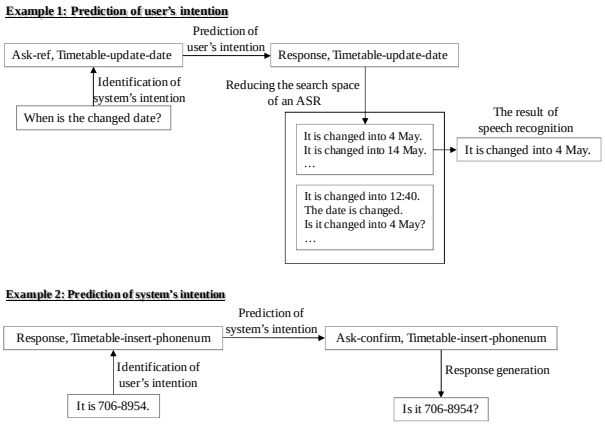

Figure 1. Motivational example

In Figure 1, the first example shows that an intention prediction module can be used as a pre-processor for reducing the search space of an ASR (automatic speech recognizer). The second example shows that an intention prediction module can be used as a pre-processor for generating a proper sentence based on dialogue history.

There are some researches on user's intention prediction (Ronnie, 1995; Reithinger, 1995). Reithinger's model used *n*-grams of speech acts as input features. Reithinger showed that his model can reduce the searching complexity of an ASR to 19~60%. However, his model did not achieve good performances because the input features were not rich enough to predict next speech acts. The researches on system's intention prediction have been treated as a part of researches on dialogue models such as a finite-state model, a frame-based

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 229–232,
Columbus, Ohio, USA, June 2008.

model (Goddeau, 1996), and a plan-based model (Litman, 1987). However, a finite-state model has a weak point that dialogue flows should be predefined. Although a plan-based model can manage complex dialogue phenomena using plan inference, a plan-based model is not easy to be applied to the real world applications because it is difficult to maintain plan recipes. In this paper, we propose a statistical model to reliably predict both user's intention and system's intention in a schedule management domain. The proposed model determines speakers' intentions by using various levels of linguistic features such as clue words, previous intentions, and a current state of a domain frame.

2 Statistical prediction of speakers' intentions

2.1 Generalization of speakers' intentions

In a goal-oriented dialogue, speaker's intention can be represented by a semantic form that consists of a speech act and a concept sequence (Levin, 2003). In the semantic form, the speech act represents the general intention expressed in an utterance, and the concept sequence captures the semantic focus of the utterance.

Table 1. Speech acts and their meanings

Speech act	Description
Greeting	The opening greeting of a dialogue
Expressive	The closing greeting of a dialogue
Opening	Sentences for opening a goal-oriented dialogue
Ask-ref	WH-questions
Ask-if	YN-questions
Response	Responses of questions or requesting actions
Request	Declarative sentences for requesting actions
Ask-confirm	Questions for confirming the previous actions
Confirm	Reponses of ask-confirm
Inform	Declarative sentences for giving some information
Accept	Agreement

Table 2. Basic concepts in a schedule management domain.

Table name	Operation name	Field name
Timetable	Insert, Delete, Select, Update	Agent, Date, Day-of-week, Time, Person, Place
Alarm	Insert, Delete, Select, Update	Date, Time

Based on these assumptions, we define 11 domain-independent speech acts, as shown in Table 1, and 53 domain-dependent concept sequences according to a three-layer annotation scheme (*i.e.* Fully connecting basic concepts with bar symbols) (Kim, 2007) based on Table 2. Then, we generalize speaker's intention into a pair of a speech act and a concept sequence. In the remains of this paper, we call a pair of a speech act and a concept sequence) an intention.

2.2 Intention prediction model

Given n utterances $U_{1,n}$ in a dialogue, let SI_{n+1} denote speaker's intention of the $n+1$th utterance. Then, the intention prediction model can be formally defined as the following equation:

$$P(SI_{n+1} \mid U_{1,n}) \approx \arg\max_{SA_{n+1}, CS_{n+1}} P(SA_{n+1}, CS_{n+1} \mid U_{1,n}) \quad (1)$$

In Equation (1), SA_{n+1} and CS_{n+1} are the speech act and the concept sequence of the $n+1$th utterance, respectively. Based on the assumption that the concept sequences are independent of the speech acts, we can rewrite Equation (1) as Equation (2).

$$P(SI_{n+1} \mid U_{1,n}) \approx \arg\max_{SA_{n+1}, CS_{n+1}} P(SA_{n+1} \mid U_{1,n}) P(CS_{n+1} \mid U_{1,n}) \quad (2)$$

In Equation (2), it is impossible to directly compute $P(SA_{n+1} \mid U_{1,n})$ and $P(CS_{n+1} \mid U_{1,n})$ because a speaker expresses identical contents with various surface forms of n sentences according to a personal linguistic sense in a real dialogue. To overcome this problem, we assume that n utterances in a dialogue can be generalized by a set of linguistic features containing various observations from the first utterance to the nth utterance. Therefore, we simplify Equation (2) by using a linguistic feature set FS_{n+1} (a set of features that are accumulated from the first utterance to nth utterance) for predicting the $n+1$th intention, as shown in Equation (3).

$$P(SI_{n+1} \mid U_{1,n}) \approx \arg\max_{SA_{n+1}, CS_{n+1}} P(SA_{n+1} \mid FS_{n+1}) P(CS_{n+1} \mid FS_{n+1}) \quad (3)$$

All terms of the right hand side in Equation (3) are represented by conditional probabilities given a various feature values. These conditional probabilities can be effectively evaluated by CRFs (conditional random fields) (Lafferty, 2001) that globally consider transition probabilities from the first ut-

terance to the *n+1*th utterance, as shown in Equation (4).

$$P_{CRF}(SA_{1,n+1} \mid FS_{1,n+1}) = \frac{1}{Z(FS_{1,n+1})} \exp\left(\sum_{i=1}^{n+1} \sum_{j} \lambda_j F_j(SA_i, FS_i)\right) \quad (4)$$

$$P_{CRF}(CS_{1,n+1} \mid FS_{1,n+1}) = \frac{1}{Z(FS_{1,n+1})} \exp\left(\sum_{i=1}^{n+1} \sum_{j} \lambda_j F_j(CS_i, FS_i)\right)$$

In Equation (4), $F_j(SA_i, FS_i)$ and $F_j(CS_i, FS_i)$ are feature functions for predicting the speech act and the concept sequence of the *i*th utterance, respectively. $Z(FS)$ is a normalization factor. The feature functions receive binary values (*i.e.* zero or one) according to absence or existence of each feature.

2.3 Multi-level features

The proposed model uses multi-level features as input values of the feature functions in Equation (4). The followings give the details of the proposed multi-level features.

- Morpheme-level feature: Sometimes a few words in a current utterance give important clues to predict an intention of a next utterance. We propose two types of morpheme-level features that are extracted from a current utterance: One is lexical features (content words annotated with parts-of-speech) and the other is POS features (part-of-speech bi-grams of all words in an utterance). To obtain the morpheme-level features, we use a conventional morphological analyzer. Then, we remove non-informative feature values by using a well-known χ^2 statistic because the previous works in document classification have shown that effective feature selection can increase precisions (Yang, 1997).
- Discourse-level feature: An intention of a current utterance affects that dialogue participants determine intentions of next utterances because a dialogue consists of utterances that are sequentially associated with each other. We propose discourse-level features (bigrams of speakers' intentions; a pair of a current intention and a next intention) that are extracted from a sequence of utterances in a current dialogue.
- Domain knowledge-level feature: In a goal-oriented dialogue, dialogue participants accomplish a given task by using shared domain knowledge. Since a frame-based model is more flexible than a finite-state model and is more easy-implementable than a plan-based model, we adopt the frame-based model in order to describe domain knowledge. We propose two types of domain knowledge-level features; slot-modification features and slot-retrieval features. The slot-modification features represent which slots are filled with suitable items, and the slot-retrieval features represent which slots are looked up. The slot-modification features and the slot-retrieval features are represented by binary notation. In the slot-modification features, '1' means that the slot is filled with a proper item, and '0' means that the slot is empty. In the slot-retrieval features, '1' means that the slot is looked up one or more times. To obtain domain knowledge-level features, we predefined speakers' intentions associated with slot modification (*e.g.* 'response & timetable-update-date') and slot retrieval (*e.g.* 'request & timetable-select-date'), respectively. Then, we automatically generated domain knowledge-level features by looking up the predefined intentions at each dialogue step.

3 Evaluation

3.1 Data sets and experimental settings

We collected a Korean dialogue corpus simulated in a schedule management domain such as appointment scheduling and alarm setting. The dialogue corpus consists of 956 dialogues, 21,336 utterances (22.3 utterances per dialogue). Each utterance in dialogues was manually annotated with speech acts and concept sequences. The manual tagging of speech acts and concept sequences was done by five graduate students with the knowledge of a dialogue analysis and post-processed by a student in a doctoral course for consistency. To experiment the proposed model, we divided the annotated messages into the training corpus and the testing corpus by a ratio of four (764 dialogues) to one (192 dialogues). Then, we performed 5-fold cross validation. We used training factors of CRFs as L-BGFS and Gaussian Prior.

3.2 Experimental results

Table 3 and Table 4 show the accuracies of the proposed model in speech act prediction and concept sequence prediction, respectively.

Table 3. The accuracies of speech act prediction

Features	Accuracy-S (%)	Accuracy-U (%)
Morpheme-level features	76.51	72.01
Discourse-level features	87.31	72.80
Domain knowledge-level feature	63.44	49.03
All features	88.11	76.25

Table 4. The accuracies of concept sequence prediction

Features	Accuracy-S (%)	Accuracy-U (%)
Morpheme-level features	66.35	59.40
Discourse-level features	86.56	62.62
Domain knowledge-level feature	37.68	49.03
All features	87.19	64.21

In Table 3 and Table 4, *Accuracy-S* means the accuracy of system's intention prediction, and *Accuracy-U* means the accuracy of user's intention prediction. Based on these experimental results, we found that multi-level features include different types of information and cooperation of the multi-level features brings synergy effect. We also found the degree of feature importance in intention prediction (*i.e.* discourse level features > morpheme-level features > domain knowledge-level features).

To evaluate the proposed model, we compare the accuracies of the proposed model with those of Reithinger's model (Reithinger, 1995) by using the same training and test corpus, as shown in Table 5.

Table 5. The comparison of accuracies

Speaker	Type	Reithinger's model	The proposed model
System	Speech act	43.37	88.11
	Concept sequence	68.06	87.19
User	Speech act	37.59	76.25
	Concept sequence	49.48	64.21

As shown in Table 5, the proposed model outperformed Reithinger's model in all kinds of predictions. We think that the differences between accuracies were mainly caused by input features: The proposed model showed similar accuracies to Reithinger's model when it used only domain knowledge-level features.

4 Conclusion

We proposed a statistical prediction model of speakers' intentions using multi-level features. The model uses three levels (a morpheme level, a discourse level, and a domain knowledge level) of features as input features of the statistical model based on CRFs. In the experiments, the proposed model showed better performances than the previous model. Based on the experiments, we found that the proposed multi-level features are very effective in speaker's intention prediction.

Acknowledgments

This research (paper) was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

References

D. Goddeau, H. Meng, J. Polifroni, S. Seneff, and S. Busayapongchai. 1996. "A Form-Based Dialogue Manager for Spoken Language Applications", *Proceedings of International Conference on Spoken Language Processing*, 701-704.

D. Litman and J. Allen. 1987. *A Plan Recognition Model for Subdialogues in Conversations*, Cognitive Science, 11:163-200.

H. Kim. 2007. *A Dialogue-based NLIDB System in a Schedule Management Domain: About the method to Find User's Intentions*, Lecture Notes in Computer Science, 4362:869-877.

J. Lafferty, A. McCallum, and F. Pereira. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting And Labeling Sequence Data", *Proceedings of ICML*, 282-289.

L. Levin, C. Langley, A. Lavie, D. Gates, D. Wallace, and K. Peterson. 2003. "Domain Specific Speech Acts for Spoken Language Translation", *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*.

N. Reithinger and E. Maier. 1995. "Utilizing Statistical Dialog Act Processing in VerbMobil", *Proceedings of ACL*, 116-121.

R. W. Smith and D. R. Hipp, 1995. *Spoken Natural Language Dialogue Systems: A Practical Approach*, Oxford University Press.

Y. Yang and J. Pedersen. 1997. "A Comparative Study on Feature Selection in Text Categorization", *Proceedings of the 14th International Conference on Machine Learning*.