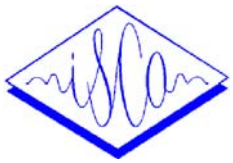
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A Robust Understanding Model for Spoken Dialogues

Junyan Chen, Ji Wu, Zuoying Wang

Department of Electronic Engineering
Tsinghua University, Beijing, China
{cjy,wuji}@thsp.ee.tsinghua.edu.cn, wzy-dee@mail.tsinghua.edu.cn

10.21437/Interspeech.2004-225

Abstract

This paper presents a robust model for language understanding in spoken dialogue systems, in which the understanding problem is formulated as a three-stage process. In the first stage, semantic concepts included in the utterance are identified through a bottom-up chart parser to build a concept graph. Then, in the second stage, communicative goal of each candidate path searched from the concept graph is determined by a classifier based on latent semantic analysis. Finally, information of discourse history is introduced to guide the search in all hypotheses for the best result. Experiments with a test set composed of spontaneous utterances show that this model outperforms a rule-based model by 8.1% and 21.5% in goal and concept understanding respectively, which proves its advantage to robust spontaneous language understanding.

1. Introduction

Although great progress has been achieved in spoken dialogue system technologies during the last few years, there are still many obstacles to be overcome before it can work efficiently in real world. Among these problems, robust spoken language understanding is rather important. Loose grammar structure of the spoken language and speech recognition errors make it quite difficult to achieve the correct understanding.

Early works in robust understanding focus on "key phrase spotting" and utterance interpretation is represented as a frame with some of its slots filled [1,2]. Here the problem of frame choosing is ignored, which may cause trouble when two utterances that contain the same key phrases have different communicative goals. The problem is especially serious in such domains with dozens of communicative goals as our train information-seeking application. Recently, more attention has been paid on communicative goal inference, such as Bayesian Belief Networks described in [3] and Maximum Entropy goal classifier proposed in [4]. Although their works have provided many valuable suggests, they didn't explicitly explain how to achieve a full understanding result with both communicative goal and semantic concepts.

Our goal is to build a robust understanding model that can obtain both the semantic concepts and the communicative goal of an utterance in our train information-seeking domain. As indicated by linguistic research, spontaneous language phenomenon is often characterized by flexible word order and frequent insertion of hesitations and speech repairs. The structure of key phrases (called as concept in this paper), however, is relatively stable and grammatical. Therefore, we decide to extract concepts through a context-free grammar in the first stage and leave the spontaneous problem to the stage of goal determination. In the second stage, a latent semantic analysis (LSA) classifier [5] is used for communicative goal determination. There are two main reasons for choosing the LSA classifier: 1) LSA doesn't care about the order of components in the utterance and is tolerable to noise, which is well-suited for spontaneous language. 2) As concept building in the first stage helps to overcome the LSA's inherent inability to capitalize on local constrains, LSA classifier's potential can be fully explored. Finally, since sentence meaning is closely related to the discourse context, we introduce the discourse information in the third stage to guide the selection of the optimal result.

The paper is organized as follows. In Section 2, we describe the three-stage understanding model from a statistical view. Then in Section 3, we proceed to a detailed discussion of each stage of the understanding process. Followed by Section 4, experimental results in our train information-seeking domain are presented and analyzed. Finally, conclusions are provided and future work discussed in Section 5.

2. Model structure

Similar to [6], the task of language understanding can be viewed as searching among all possible discourse-dependent meanings M_D for the most likely meaning M_D^* given the input word string W and dialogue history H:

$$M_D^* = \arg\max_{M_D} P(M_D \mid W, H) \quad (1)$$

To simplify the problem, discourse-independent meaning M_S is introduced as an intermediate level, which gives:

$$M_D^* = \arg\max_{M_D} \sum_{M_S} P(M_D \mid M_S, W, H) P(M_S \mid W, H) \quad (2)$$

where M_S is independent of the dialogue history and can be viewed as composed of two parts: communicative goal G_S and semantic concepts C_S of the utterance. Therefore, equation (2) can be rewritten as:

$$M_D^* = \arg\max_{M_D} \sum_{M_S} P(M_D \mid M_S, W, H) P(G_S, C_S \mid W) \quad (3)$$

After introducing Bayesian rule and assuming that M_D is independent of W when M_S and H are given, equation (3) can be rewritten as:

$$M_D^* = \arg\max_{M_D} \sum_{M_S} P(M_D \mid M_S, H) P(G_S, C_S, W) \quad (4)$$

In equation (4), denominator $P(W)$ is omitted since it is constant for all M_D and M_S. Now, if we assume G_S is

independent of W once the semantic concepts C_S in the utterance are determined, then:

$$M_D^* = \arg\max_{M_D} \sum_{M_S} P(M_D \mid M_S, H) P(G_S \mid C_S) P(C_S, W) \quad (5)$$

Thus the understanding task can be viewed as a three-stage process:

1) Concept understanding based on $P(C_S, W)$;
2) Goal determination based on $P(G_S \mid C_S)$;
3) Discourse-dependent meaning understanding based on $P(M_D \mid M_S, H)$.

In practice, as it is difficult to take into account of all possible M_S of an utterance, we choose to search for the optimal result from an n-best path list. This is reasonable since most of the probability mass is focused on the first n-best candidates.

3. The three-stage understanding

3.1. Concept identification

The task of this stage is to extract concepts from the recognition result. Due to the relatively stable structure of the concepts, we manually designed a set of context-free rules for each of the 41 concept types (such as train number, date, destination, etc.) in our train information-seeking domain. A bottom-up chart parser is used to bind words up into concept phrases according to these rules. To avoid eliminating the correct result too early, word graph output by the speech recognizer is used as the input to the understanding module. Each time a word graph is received, the bottom-up chart parser searches it for concept phrases and stores them in a concept graph. Words that could not be bound up by any rule are viewed as concepts themselves and are added to the concept graph directly. After the concept graph is created, n-best candidate paths are searched from the graph according to the acoustic score and the model score from $P(C_S, W)$. To simplify the problem, the model is further decompose as

$$P(C_S, W) = \prod_{i=1}^{N} P(C_i \mid C_{i-1}) \prod_{j=1}^{M} P(w_j \mid w_{j-1}, C_i) \quad (6)$$

In equation (6), the following two approximations are used:

$$P(C_i \mid C_1^{i-1}, W_1^{N_{C_{i-1}}}) \approx P(C_i \mid C_{i-1}) \quad (7)$$

$$P(w_j \mid W_1^{j-1}, C_1^i) \approx P(w_j \mid w_{j-1}, C_i) \quad (8)$$

where $C_S = (C_1, C_2, \ldots, C_N)$, $W = (w_1, w_2, \ldots, w_M)$, W_1^j denotes the word string $w_1, w_2, \ldots, w_j$ and similarly for concepts. $W_1^{N_{C_i}}$ is the word string corresponding to the concept string $(C_1, C_2, \ldots, C_i)$. Henceforth, path score in this stage can be calculated as the sum of acoustic score, concept bigram score and the concept-depended word bigram score.

Since our current data from actual dialogues are very limited, we hand-designed a set of sentence level context-free rules and generated about 100,000 user utterances to train the two bigram models. To alleviate the influence of sparse data, both the concept bigram and the concept-depended word bigram are smoothed with the Kneser-Ney approach [7].

3.2. LSA-based goal determination

3.2.1. Building the LSA co-occurrence matrix

In this stage, a latent semantic analysis classifier is used for communicative goal determination for each candidate path provided in the first stage. First, we construct a unit-goal co-occurrence matrix in a manner similar to that detailed in [8]. The row variables correspond to all possible communicative goals in our train information-seeking domain and the column variables correspond to a special-defined unit set. It is composed of two types of unit: concept and free word. Free words are those that can occur as concept themselves in concept graphs, such as "of", "that", etc. Words belonging to some basic semantic types are not free. For example, the word "Beijing" will be identified as the semantic type "CITY" once occurred in the word graph. There are 19 communicative goals and 778 units in our current system and thus the co-occurrence matrix is 778×19.

Corpus used to train the matrix is the same as that for training the two bigram models in the first stage. Because all utterances are generated from rules, it leads to a LSA classifier more accurate for utterance in accordance with the hand-designed rules, which is proved by experiments in Section 4.2.

3.2.2. Communicative goal determination

The determination of the communicative goal of each candidate path followed the idea in [5]. After comparing the cosine distance between each goal and the variant vector corresponding to the current path, the goal with the nearest distance is chosen with classification score added to the path score. The classification score is defined as the negative logarithm of $P(G_S \mid C_S)$, and

$$P(G_S \mid C_S) = \frac{dist(g_i^*, v)}{\sum_{g_i} dist(g_i, v)} \quad (9)$$

where v denotes the variable vector of the path, g_i represents communicative goal of the domain, g_i^* is the goal that has the nearest distance from v.

Since the variable vectors of short utterances are too sparse to be reliable, we only perform LSA goal classification for utterances longer than a threshold and use a rule-based robust parser described in Section 4.1 for short ones.

3.3. Discourse-dependent meaning understanding

In this stage, dialogue history is introduced to pick out the most probable discourse-dependent meaning M_D^* according to $P(M_D \mid M_S, H)$, which can be approximated by:

$$\begin{aligned} P(M_D \mid M_S, H) &= P(G_D, C_D \mid G_S, C_S, H) \\ &= P(G_D \mid G_S, C_S, H) P(C_D \mid C_S, G_S, G_D, H) \end{aligned} \quad (10)$$

If the following two independent assumptions are made:

1) When dialogue history and discourse-independent goal are given, discourse-dependent goal does not depend on concepts, that is

$$P(G_D \mid G_S, C_S, H) \approx P(G_D \mid G_S, H) \quad (11)$$

2) Dialogue history does not affect the goal of the utterance determined in early stages, which means:

$$P(G_D \mid G_S, H) \approx P(G_S \mid H) \quad (12)$$

$$P(C_D \mid C_S, G_S, G_D, H) \approx P(C_D \mid C_S, G_S, H) \quad (13)$$

Then, equation (10) can be simplified as:

$$P(M_D \mid M_S, H) \approx P(G_S \mid H) P(C_D \mid C_S, G_S, H) \quad (14)$$

For $P(G_S \mid H)$, information of user utterance expectation provided by a user plan inference model detailed in [9] is used. At each turn of the dialogue, the inferencer estimates what the user would say based on a user plan model and assigns each goal g_i a probability p_i.

For $P(C_D \mid C_S, G_S, H)$, we introduce a set of consistency constraints. They are within-interpretation constraint, history-depended constraint and common sense constraint, which are listed below:

1. Within-interpretation constraint: If contradictory values for a semantic concept are included in the path, a penalty score $Penalty_1$ is added to the path score.
2. History-depended constraint: If semantic concept included in the path contradicts with the user plan estimated by the plan inferencer, a penalty score $Penalty_2$ is added to the path score.
3. Common sense constraint: If semantic concept included in the path contradicts with some common sense, a penalty score $Penalty_3$ is added to the path score. (For example, that the departure station should not be the same with the arrival station is a common sense in the train information-seeking domain.)

Therefore, score of $P(M_D \mid M_S, H)$ can be calculated as:

$$\begin{aligned} &-\log(P(M_D \mid M_S, H)) \\ &\quad \approx -\log(P(G_S \mid H) P(C_D \mid C_S, G_S, H)) \\ &\quad = -\log(p_i) + \sum_i Penalty_i \end{aligned} \quad (15)$$

After adding this score to each candidate path, the path with the smallest score is selected with its communicative goal and semantic concepts taken as the final understanding result of the current user utterance.

4. Experiments

4.1. Comparing with a rule-based model

To see how well the current three-stage understanding model performs, we implemented a rule-based understanding module for benchmarking purposes. In the baseline system, the first step is the same, which generates a concept graph from the word graph output by the speech recognizer. Then a robust top-down chart parser searches the concept graph for a best path based on the sentence level rules that have been used to generate the training corpus. During the top-down parsing, filler arcs are introduced to skip unimportant words. User plan inference information is also integrated to dynamically change the rules and their corresponding weights for understanding the user input of the next turn.

Two test sets are selected for carrying out the comparing experiments. The first one contains 118 dialogues, 512 user utterances, all of which comply with the sentence level rules. The second test set is from transcriptions of 84 actual user-system dialogues, containing 494 user utterances with a lot of ungrammatical and spontaneous phenomena. For either test set, we had 8 people record the utterances. Speech data from these 16 speakers are fed to the speech recognizer to generate the corresponding word graphs, which are used as inputs to the understanding module. Besides, since current experiments are conducted without integrating the dialogue manager, correct user plan inference information for each utterance are manually prepared to guide the understanding.

Understanding errors of the baseline system and the three-stage system on both test sets are presented in Table 1, where INS, DEL and SUB denote concept insertion, deletion and substitution error rate. CER, the total concept error rate, is the sum of INS, DEL and SUB. GER is the goal determination error rate.

Table 1: Understanding error rate of the three-stage understanding model versus a rule-based robust understanding model on test set one

		INS (%)	DEL (%)	SUB (%)	CER (%)	GER (%)
Test Set 1	Baseline	12.3	15.7	8.7	36.7	24.5
	Three-stage	8.3	10.6	5.7	24.6	26.0
	Relative error reduction	32.5	32.5	34.5	33.0	-6.1
Test Set 2	Baseline	28.8	41.3	12.5	82.6	49.6
	Three-stage	21.5	33.7	9.6	64.8	45.6
	Relative error reduction	25.3	18.4	23.2	21.5	8.1

It is observed that on the first set, the three-stage understanding model achieves much higher accuracy in concept understanding compared to the rule-based model, with CER decreased by 33%. A possible explanation may be that in the three-stage model constraints among concepts are much more loose than that of the rule-based model. Although there is a slight degradation in communicative goal determination, it is acceptable because the rule-based understanding model should do well in grasping the meaning of grammatical utterances.

What encourages us is the great improvement on the second set. Relative error reduction to the rule-based model achieves 21.5% and 8.1% for concept and communicative goal understanding respectively. Since the second set is composed of actual utterances with many spontaneous

phenomena, improvement on this set implies that the three-stage model is more robust at handling spoken language.

4.2. Recognition errors on performance

The purpose of this experiment is to gauge the gap in performance caused by speech recognition errors. Transcriptions for utterances in both test sets that are long enough to be processed in LSA mode are fed to the three-stage understanding system. Recognition errors of the n-best (n=100 in current system) candidate paths are calculated accordingly. Results for both sets are tabulated in Table 2 where INS, DEL, SUB and CER denote word insertion, deletion, substitution and total error rate for the line of "Recognition errors". For comparison, the results of the three-stage model are listed again.

Table 2: Transcription understanding errors versus speech recognition errors

		INS (%)	DEL (%)	SUB (%)	CER (%)	GER (%)
Test Set 1	Transcript	0	0	0	0	7.5
	Three-stage	8.3	10.6	5.7	24.6	26.0
	Recognition errors	2.1	4.9	13.5	20.5	—
Test Set 2	Transcript	0	1.2	0.7	1.9	17.3
	Three-stage	21.5	33.7	9.6	64.8	45.6
	Recognition errors	1.1	7.5	28.4	37.0	—

As shown in Table 2, the recognition error rate does have a substantial impact on the understanding performance, especially on concept error rate. This is reasonable since concept understanding is mainly rule-based and thus vulnerable to recognition errors. Goal determination based on the statistical LSA, however, is more robust and therefore degrades more gracefully.

Another issue to note is that the goal understanding accuracy for transcriptions is much worse on the second set than on the first one. One possible reason is that the spontaneous input really makes the problem more difficult. But we think a more important factor is that the lack of training data from actual human-machine dialogues. As mentioned in Section 3, all training data are rule-generated, which inevitably causes the LSA co-occurrence matrix poor-adapted to spontaneous utterances that appear frequently in the second set. We believe that the performance on spontaneous utterances will be improved if more data collected from actual dialogues are used for training.

Analysis of the understanding errors shows that some goals are inherently close in the LSA space, which makes the classification difficult. However, it can be improved if we refine our current simple discourse model by removing the assumption that dialogue history does not affect the goal of the utterance determined in early stages. To illustrate as an example, consider the dialogue in Figure 1. As can be seen, the second user utterance is asking whether some other types of ticket exist. The LSA classifier, however, does not know this and will be in a dilemma of choosing between the goals of "QueryTicketExist" and "QueryTrainExist" in our current system. However, if the dialogue history is considered in this case, it is easy to make the right choice. Examination of the goal determination errors indicates that about 25% errors can be avoided if discourse information is made full use of.

User: Em, is there any ticket for sleeping car of train No.21 on Friday?
System: Sorry, tickets for sleeping car of train No.21 on Friday have been sold out.What else can I do for you?
User: Is there any other types?
... ...

Figure 1 Example illustrating context-based understanding

5. Conclusions

In this paper, we propose a robust understanding model in which the understanding process is accomplished through three stages: concept identification, LSA-based goal determination and discourse-dependent meaning understanding. Loose constraints on concepts relation and sentence structure make the three-stage model more robust to recognition errors and spontaneous language than a rule-based model. Experiments on a test set composed of utterances from actual dialogues indicate that it outperforms the rule-based model by 8.1% and 21.5% in goal and concept understanding respectively.

Future work includes using data from actual dialogues for training to improve the system performance on spontaneous utterances, as well as refining the discourse model for more accurate goal understanding.

6. References

[1] Ward, W., "Understanding Spontaneous Speech: the Phoenix System", in *Proc. of ICASSP*, 1991, pp.365-367.

[2] Kellner, A., Rueber, B., Seide, F., and Tran, B. H., "PADIS – An Automatic Telephone Switchboard and Directory Information System", *Speech Communication*, 23(2): 95-111, 1997.

[3] Meng, H. M., Lam, W. and Wai, C. "To Believe is to Understand", in *Proc EuroSpeech*, 1999, pp. 2015-2018.

[4] Chelba, C., Mahajan, M. and Acero, A., "Speech Utterance Classification", in *Proc. of ICASSP*, 2003, pp.280-283.

[5] Bellegarda, J. R., Silverman, K. E. A., "Natural Language Spoken Interface Control Using Data-Driven Semantic Inference", *IEEE Trans. Speech and Audio Proc.*, 11(3): 267-277, 2003.

[6] Miller, S., Stallard, D., Bobrow, R. and Schwartz, R. "A Fully Statistical Approach to Natural Language Interfaces", in *Proc. of ACL*, 1996, pp.55-61.

[7] Chen, S. F. and Goodman, J., "An Empirical Study of Smoothing Techniques for Language Modeling", *Computer Speech and Language*, 13(4): 359-394, 1999.

[8] Bellegarda, J. R., "Exploiting Latent Semantic Information in Statistical Language Modeling", *Proceedings of the IEEE*, 88(8): 1279-1296, 2000.

[9] Chen, Junyan, Wu, Ji and Wang, Zuoying, "A Chinese Spoken Dialogue System for Train Information", in *Proc. of IEEE SMC*, 2003, pp.1097-1103.