

Extracting Domain Knowledge for Dialogue Model Adaptation

Kuei-Kuang Lin and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
hh_chen@csie.ntu.edu.tw

Abstract. Domain shift is a challenging issue in dialogue management. This paper shows how to extract domain knowledge for dialogue model adaptation. The basic semantic concepts are derived from domain corpus by iterative token combination and contextual clustering. Speech act is identified by using semantic clues within an utterance. Frame states summarize current dialogue condition and state transition captures the mental agreement between users and system. Both Bayesian and machine learning approaches are experimented in identification of speech act and prediction of next state. To test the feasibility of this model adaptation approach, four corpora from domains of hospital registration service, telephone inquiring service, railway information service and air traveling information service are adopted. The experimental results demonstrate good portability in different domains.

1 Introduction

Dialogue management provides a rich human-computer interaction which allows users to convey more complex information than a single utterance. Despite of the recent significant progress in the areas of human language processing, building successful dialogue systems still requires large amounts of development time and human expertise [1]. The major challenging issue is the introduction of the new domain knowledge to the dialogue model when domain is shifted. That usually takes time to handcraft the domain knowledge that a dialogue manager needs. In the past, some papers [2,6] dealt with acquisition and clustering of grammatical fragments for natural language understanding; and some papers [4,9] employed statistical techniques for recognizing speech intentions. This paper emphasizes on how to extract crucial domain knowledge, including semantic concept extraction, speech act identification and formulation of dialogue state transition. Four corpora from different domains are employed to test the feasibility.

2 Corpora of Different Domains

Two dialogue corpora and two single query corpora were studied. They belong to domains of hospital registration service, telephone inquiring service, railway information service and air traveling information service. Tables 1 and 2 summarize the statistics of the materials. The dialogues in NTUH corpus, which were transcribed from face to face conversation in Chinese, deal with tasks in a registration counter of NTU hospital, including registration, cancellation, information seeking, *etc.* The Chinese CHT corpus was transcribed from Chun-Hwa Telecom phone number inquiring system through telephone. Compared with NTUH corpus, most of the utterances in CHT corpus are very short and incomplete due to the fact that people often address the targets directly when using phone number inquiring service.

Table 1. Dialogue Corpora

Corpus name	NTU Hospital (NTUH)	Chun-Hwa Telecom (CHT)
Content	Register (make appointment), cancel, inquire info	Inquire phone number or other info
Number of dialogues	13	98
Number of utterances	440	1923
Average length	33.5	16.4

Besides the dialogue corpora, two Chinese query corpora, Taiwan Railway Corpus (TWR) and Air Traveling Information Service corpus (CATIS), which include queries about train timetable and air traveling information, respectively, were employed. CATIS is a Chinese version of ATIS [7]. All the airline booking information, e.g., location names, airline names, *etc.*, are translated into Chinese. CATIS is much larger, and it contains more unknown words than TWR corpus.

Table 2. Corpora of Single Utterances

Corpus name	Taiwan Railway (TWR)	Air Traveling (CATIS)
Content	Train timetable queries	Airline booking queries
Number of utterances	200	5517
Average length	29	32

3 Acquisition of Semantic Concepts

Semantic concept refers to key entities in a domain that users have to fill, answer or mention to accomplish the desired tasks. Concepts come with different forms, e.g., they could be database attributes, certain key verbs, or some types of named entities. A concept may have several values, e.g., a destination station in railway corpus may be any location names. In the proposed data-driven methodology, token combination is performed first to combine tokens, and then contextual cluster is employed to

gather terms with similar context. The combined tokens are labeled to create a modified corpus for another iteration of token combination and contextual clustering.

NTUH dialogue corpus is adopted for experiment. The Chinese corpus is segmented and tagged with parts of speech. Unseen words like person names are often the major source of segmentation errors. For example, a doctor's name “楊士毅” (Yang Shi Yi) was segmented to three individual characters. Named entity recognition [3] attempts to merge such individual characters. Besides named entities, certain word strings denote potential semantic concepts. We group terms that tend to co-occur in NTUH corpus by mutual information shown as follows. They form phrases or multi-word entities.

$$MI(e_1, e_2) = P(e_1|e_2) \log \frac{P(e_1, e_2)}{P(e_1)P(e_2)} \quad (1)$$

Terms of the same semantic concepts are often represented with similar utterance structure and neighbor context. For example, “我要查第一銀行的電話” (I want the phone number of First Bank) and “我要查台灣大學的電話” (I want the telephone number of National Taiwan University) are used in CHT corpus. The two target elements “第一銀行” (First Bank) and “台灣大學” (National Taiwan University) have similar left and right contexts, which show that two different names denote the similar concepts. Kullback-Leibler distance [5] is used to measure the similarity between two contexts, where V denotes the vocabulary used in contexts.

$$D(p_1 \| p_2) = \sum_{i=1}^V p_1(i) \log \frac{p_1(i)}{p_2(i)} \quad (2)$$

$D(p_1 \| p_2) = 0$ if p_1 and p_2 are equivalent, i.e., they have exactly the same neighboring context. Terms with large MI are merged into a larger term and terms with small KL distance are grouped into the same cluster.

Experimental results are judged by human assessors. The results of token combination are rated as four levels – say, *correct word*, *correct phrase*, *nonsense* (i.e., wrong combination) and *longer* (i.e., terms contains both correct and incorrect combinations of previous iterations). Figure 1 shows that formulation of *words* starts from the beginning of token combination, and grows steadily until middle of process (around iteration 30). Phrase combinations occur later at the 15th iteration and the number does not increase as fast as word combinations do. Because the number of phrases is smaller and the meaningful combinations are formulated from word level to phrase level, the number of nonsense and longer combinations increases rapidly in the later iterations. Error propagations exist after 10-15 iterations and result in most nonsense combinations as the curve goes approximately along with nonsense curve.

Contextual clusters are rated as three levels – say, *correct* (i.e., all the clustered terms belong to the same semantic concept), *wrong* (i.e., all the terms are unrelated to one another), and *part* (i.e., some of the clustered items belong to same concept, and some are not.) Figure 2 shows that the curves of correct (meaningful) and wrong (meaningless) clusters have same tendency. They all increase rapidly in the first 5 iterations, and then the increase speed slows down. As the iteration proceeds, a cluster containing totally unrelated terms would seldom occur because terms with clear evidence have been correctly or partly clustered in the previous iterations. Error

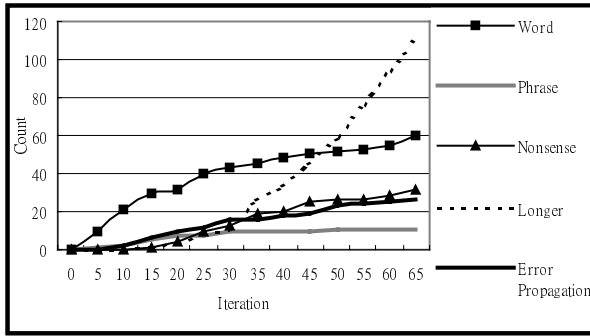


Fig. 1. Quantitative Result of Token Combination

propagation begins at the 10th iteration and grows with the number of “part” clusters. The number of “part” clusters stops growing after 30 iterations. These clusters are in fact unrecognizable for human to make judgment. We can see that the meaningful clusters are proposed in the first half of iterations, and most of the later iterations provide useless clusters. Therefore a reasonable number of iterations should be inspected to stop the clustering algorithm.

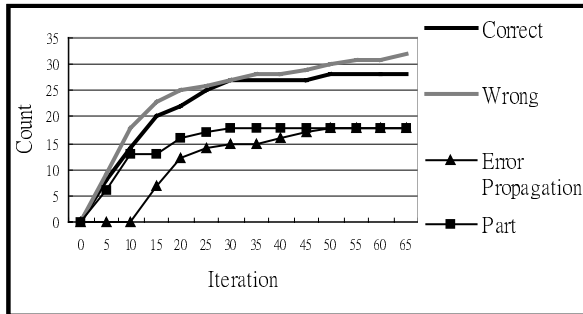


Fig. 2. Quantitative Result of Contextual Clustering

A concept category groups semantic concepts that serve similar roles or present similar intentions in an utterance. At this time, the meanings of the clusters generated from previous experiments are undefined. After the clustering results are examined, total 44 concepts are identified. We re-label the original corpus with these concepts and cluster the corpus again. Table 3 shows some examples of the result.

Although the number of meaningful categories is only few, the result indeed indicates that some concepts can be grouped to proper categories. A complete concept set is formulated by using the proposed semi-automatic algorithm. The experiments show the algorithm is helpful for human to craft domain knowledge. The extracted semantic concepts will be used in the following sections.

Table 3. Categorization of Semantic Concepts

Concept Category	Semantic Concept
Domain Slot Values	(BIRTHDATE_VALUE), (IDNUMBER_VALUE), (RECORDNUM_VALUE), (DATE_VALUE), (TIME_VALUE)
Domain Slot Names	(DATA), (ID_NUM), (RECORD_NUM), (PATIENT_TYPE), (DEPARTMENT_TYPE)
Domain Specific Actions	(CHECK), (CANCEL), (CHANGE), (REGISTER)
Frequently Used Verbs	(WANT), (REQUEST)

4 Identification of Speech Act

Speech acts represent the mental state of the participants in a dialogue. Some words in each utterance are replaced with the semantic concepts derived in last section. A set of speech acts are defined, including *Request-ref*, *Answer-ref*, *Request-if*, *Answer-if*, *Request-fact*, *Answer-fact*, *Greeting*, *Prompt*, *Clarify*, *Accept*, and *Reject* [4]. Four trained annotators were asked to tag corpora with these speech acts. That will serve as an answer set for evaluation. Formula 3 defines the Bayesian identification. Given a set of conceptual features in an utterance, which are denoted by semantic tags C_k , we try to find the speech act with the largest probability. Here we assume that each concept is independent of each other.

$$P\left(\hat{A} \mid \vec{C}\right) = \arg \max_{A_i} P\left(A_i\right) \prod_{k=1}^M \frac{P\left(C_k = c_k \mid A_i\right)}{P\left(C_k = c_k\right)} \quad (3)$$

The precision of the experimental result is 57%. Compared to raw material (i.e., words with the N highest *tf*idf* in an utterance are regarded as features), which has only 18% of precision, semantic concepts eliminate data sparseness problem. The bad performance of raw data is due to the small size of the corpus from which *tf* and *idf* are trained. To tell if the derived concepts are redundant, we divide the concepts into several subgroups and make the similar experiments again. We find out that the best performance occurs when all the concepts are adopted. It shows that the derived semantic concepts do capture specific features of domain utterances and are not redundant.

See5 [8], a machine learning tool, is also adopted to identify speech act. Semantic tags extracted from each utterance are used as attributes to identify speech act. The precision is 65% when all the concepts are considered as clues. Compared to the performance of using the raw material, i.e., 53%, semantic concepts got less gain than Bayesian method.

5 Dialogue Modeling

We adopted frame-based model to formulate the dialogue behavior in this paper. The agreements of contexts between participants in a dialogue are based on the content of frame slots. Predicting the condition of each frame slot could capture the transition

process of dialogues. An information state consists of several semantic slots selected from the semantic concepts computed in Section 3. Besides, several conditions are defined to represent each slot state, including *in question*, *mentioned*, *with value*, and *empty*, which are denoted by symbols *, +, 1, and 0, respectively, in Table 4. Originally, all the frame slots are set to 0. An algorithm shown as follows determines the state of each frame slot.

1. Check if any slot value is confirmed. If so, fill the slot with value and set the state to 1.
2. Determine if any questioning clues exist. If yes, tag each slot name mentioned along with questioning clues, and set the state to *.
3. Tag all mentioned slot names as mentioned, and set the state to +.

Using this algorithm, an input utterance will be transformed into information state representation. Total 13 slots, including *branch name*, *service type*, *department type*, *doctor name*, *week date*, *time*, *birth date*, *ID number*, *medical record number*, *patient name*, *number of times coming*, and *general number*, are selected from the derived semantic concepts, and form the set of frame states. Not all slots are presented in the Table 4. Numbers 1-5 show the first five slots illustrated above.

Table 4. Examples of Transformation of Information State

Original Utterance	U: 我要掛眼科門診。(I want to register outpatient service of ophthalmology department.)	U: 不知道星期幾有呢?(I wonder on what day there is such a service?)	S: 請問要在總院看還是公館分院看?(Would you like to go to main hospital or Gong-gwan branch?)
Concepts chunked	(department_type) (service_type) (person) (want) (register)	(week) (question_word)	(branch_name1) (branch_name2) (request) (at) (see_doctor) (question_word)
Slot	User	User	System
1	0	0	+
2	1	1	1
3	1	1	1
4	0	0	0
5	0	*	*

Following the above procedure, a dialogue corpus could be transformed into transitions of a sequence of information states. By using an information state as a clue, we have transformed the problem from modeling a complete dialogue into predicting next information state. Because the history of dialogue is accumulated in terms of state transition, every prediction to the next state actually concerns about all the previous dialogue history.

Due to the sparseness and small amount of corpus, Bayesian prediction is divided into two phases: 1) Predict the entire set of states when the previous state actually occurred in the training corpus; and 2) If not, assume that each slot is independent, and predict each individually. In Formulas 4, 5, 6 and 7, S_{i+1} is the whole frame state in time $i+1$; s_{i+1} is a candidate of the next slot state; s_i is the current

slot state; f_j is a feature presented in the current utterance; and K is the number of features.

$$P(S_{i+1}|S_i) = P(S_{i+1}) \frac{P(S_i|S_{i+1})}{P(S_i)} \quad (4)$$

$$P(s_{i+1}|s_i) = P(s_{i+1}) \frac{P(s_i|s_{i+1})}{P(s_i)} \quad (5)$$

$$P(s_{i+1}|\vec{F}) = P(s_i) \prod_{j=1}^K \frac{P(F_j = f_j|s_{i+1})}{P(F_j = f_j)} \quad (6)$$

$$score(s_{i+1}) = P(s_{i+1}|s_i) * P(s_{i+1}|\vec{F}) \quad (7)$$

In the experiment, there are 440 transitions. In 209 of these transitions, the previous frame state occurred in the training corpus. The predictions are done trivially in phase 1. In the other 231 transitions, each slot state is predicted in phase 2. The next frame slots can be predicted from the training experience in about 60% (i.e., 127 correct among 209 example frames) of the transitions. If the determination is relaxed to determining if a slot state should appear, the precision is up to 82%. Table 5 summarizes the experimental results of phase 1. It shows that if there is a large enough training corpus to obtain more reliable dialogue states, the prediction of next frame state is feasible. For those not seen in the training set, phase 2 predicts each slot state respectively. Table 6 shows the result. Each frame contains 13 slots to be predicted. The overall precision is 91.2%. After further analyzing the distribution of slot state, we found that most of the slot states (i.e., 97.4%) are not change through transitions. In other words, it is much easier to predict an unchanged slot state than a changed one. In our experiment, 50.6% of slot states which are changed can be correctly predicted, on the other hand, 92.34% of slot states which are not changed can be predicted correctly.

Table 5. Phase 1 Result

Number of frame	209
Correct	127
Plain Correct	45
Incorrect	37
Precision	82.3%

Machine learning method is also applied to predict each slot state. Semantic features, the previous slot state, the role and the speech act of current utterance are considered as the attributes for each slot state. The result shows that the most important attributes are the previous slot states. The precision drops from 98% to 7% without consideration of the previous slots.

Table 6. Phase 2 Result

Number of frame	231
Number of slots	3003
Correct	2740
Incorrect	263
Precision	91.2%

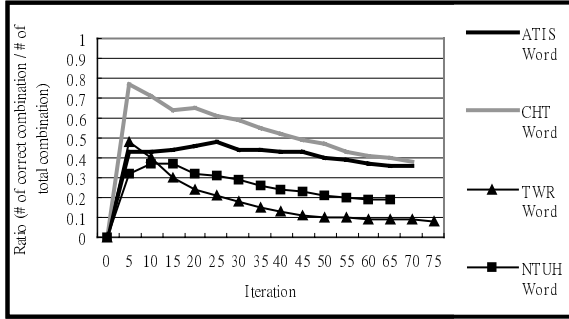


Fig. 3. Results of Word Combination in Different Corpora

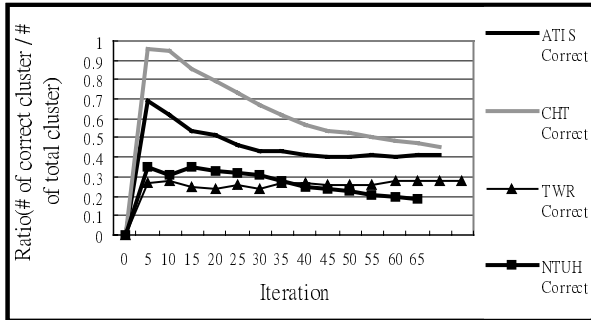


Fig. 4. Results of Clustering in Different Corpora

6 Shifting Domain

The proposed method is experimented on the NTUH service domain in the previous sections. To test its portability, the method is also applied to the other three different domains.

In semantic concept acquisition, we can see that each domain presents similar tendency, as shown in Figures 3 and 4. The difference lies mainly in different sizes and characteristics of corpora. For example, CHT corpus contains many named

entities and special representations, so that the proportion of correct combinations is higher.

In speech act identification, the precision rates in CHT corpus are 77% with machine learning method and 57% with Bayesian method. CHT is comparatively a simpler domain than NTUH. The most frequent speech acts in CHT corpus are *Request-ref* and *Answer-ref*. In dialogue modeling, average frame state prediction (phase 1) is 85%, and slot state prediction (phase 2) is 87%.

7 Conclusions and Future Work

This paper proposes a systematic procedure to extract domain knowledge for dialogue adaptation. Semantic acquisition extracts key concepts semi-automatically to decrease human intervening cost. Speech act identification recognizes current intent and focus of an utterance. Regarding derived features as frame states, dialogue transition is modeled as prediction of frame states. Applying the procedure to four different domains shows its portability. More data collection and domains will be experimented in future work.

References

1. Allen, J.F. *et al.*: Towards Conversational Human-Computer Interaction. In AI Magazine, (2001)
2. Arai, K. *et al.*: Grammar Fragment Acquisition using Syntactic and Semantic Clustering. In Speech Communication, Vol. 27, No. 1, (1999)
3. Chen, H.H., Ding, Y.W. and Tsai, S.C.: Named Entity Extraction for Information Retrieval. In Computer Processing of Oriental Languages, Vol. 12, No. 1, (1998) 75-85.
4. Chu-Carrol, J.: A Statistical Model for Discourse Act Recognition in Dialogue Interactions. In Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing (1998)
5. Kullback, S.: Information Theory and Statistics. John Wiley and Sons, New York (1959)
6. Meng, Helen and Siu, K.C.: Semiautomatic Acquisition of Semantic Structures. In IEEE Transaction Knowledge and Data Engineering, Vol. 14, (2001)
7. Price, P.: Evaluation of Spoken Language Systems: The ATIS Domain. In Proceedings of ARPA Human Language Technology Workshop (1990)
8. Quinlan, R.: See5. URL: www.rulequest.com (2002)
9. Stolcke, A. and Shriberg, E.: Dialog Act Modeling for Conversational Speech. In Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing (1998)