A Syllable-Based Chinese Spoken Dialogue System for Telephone Directory Services Primarily Trained with A Corpus

Yen-Ju Yang[1] *and Lin-Shan Lee*[1,2]

[1]Dept. of Computer Science and Information Engineering, National Taiwan University
[2]Institute of Information Science, Academia Sinica
Taipei, Taiwan, 106, ROC
e-mail: kathy@speech.ee.ntu.edu.tw

ABSTRACT

This paper presents a syllable-based Chinese spoken dialogue system for telephone directory services primarily trained with a corpus. It integrates automatic phrase extraction, robust phrase spotting, statistics-based semantic parsing by phrase-concept joint language model as well as concept-based dialogue model, and intention identification by probabilistic finite state network to form a speech intention estimator. By applying the proposed techniques, the concept sequence with the maximum a-posteriori (MAP) probability based on intra and inter sentence consideration conveyed in the user's speech sentence, i.e. the speaker's intention, can be identified. This approach is convenient to be trained by a given corpus and flexible to be ported to different dialogue tasks. Incorporate a mixed-initiative goal-oriented dialogue manager, we have successfully developed a dialogue system for telephone directory service. Very promising results have been obtained in on-line tests.

1. INTRODUCTION

Quite many successful spoken dialogue systems have been developed all over the world in recent years, and many promising applications have been identified, although not too much results have been reported on Chinese dialogue systems [1]. Chinese language is monosyllabic structure, i.e. almost every character in Chinese is a morpheme with its own meaning, and is pronounced as a monosyllable. As a result, the wording structure in Chinese is quite flexible. For example, many words can be arbitrarily abbreviated, while the system needs to be able to handle them. The syllable-based approach, in which the basic unit for recognition is the syllable rather than the word, is found very helpful, because the syllables correspond to exactly the characters with meaning.

There are more complicated phenomena in spontaneous speech such as the lower level events like pauses, filled pauses (e.g. "uh"), hesitation, laughter as well as other non-speech noises (inhalation, cough); and the higher level events like false starts, restarts, etc. In addition, recognition errors, out-of-vocabulary (unknown words) and out-of-grammar occur more in spontaneous speech than read speech. In order to deal with above problems for the natural language analysis in spoken dialogue system, almost all viable systems have abandoned the notion of achieving a complete syntactic analysis of every input sentence, favoring a more robust strategy that can still answer when a full parse is not achieved [2-4]. This can be accomplished by identifying parsable phrases and clauses, and providing a separate mechanism for grouping them together to form a complete meaning analysis. According to this concept, we present a different way called speech intention estimator to accomplish. First, we try to automatically extract domain specific phrase lexicon from corpus. During understanding phase, robust phrase spotting is applied on the syllable lattice (result of recognition phase) and then the phrase-level parsing is performed on the phrase lattice to get the complete meaning. Since a sentence is not composed of **many** phrases, the parsing process can be well handled by phrase-level n-gram language models, which integrate phrase, semantic concept and dialogue model to modeling intra and inter sentence structure. Finally, intention identification is used to classify the complete meaning to a higher-level intention abstraction via probabilistic finite state network for further processed by dialogue manager. The proposed system block diagram and the detail speech intention estimator are shown in Figure 1 and 2.

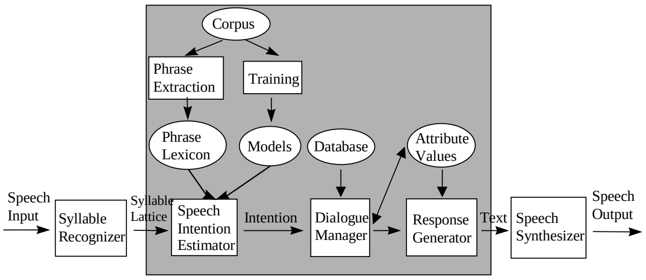

Figure 1: Block diagram of presented spoken dialogues.

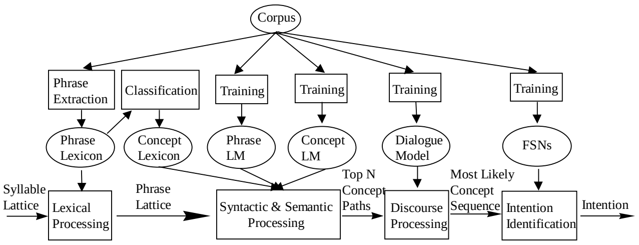

Figure 2: Speech intention estimator.

2. AUTOMATIC PHRASE EXTRACTION

The utterance persons interacting with each other in a specific task always includes several recurrent phrases, or called collocations. These phrases have their specific meanings, which can further lead to partial or full understanding. As a consequence, we think phrases are "significant and frequent cooccurred patterns relevant to domain-dependent subject." This means they have high association among components and are not all the same for different domain subject. So we try to automatically extract phrase lexicon for a certain task from dialogue corpus, which represents the combination of frequently uttered vocabulary in the conversation.

A phrase is defined as a string composed of from one to several words representing syntactic and semantic information. Due to special structure of Chinese language, each Chinese character is pronounced as a monosyllable, and a Chinese word is composed of from one to several characters. Since phrase is a combination of words, a Chinese phrase is also composed of from one to several characters. So we adopt the bottom-up strategy to extract the syllable/character patterns iteratively from characters to words and further to partial phrases until no other components have high association with them. We have presented an efficient approach to measure association [5]. The result shows it really can extract significant and recurrent phrase, but some false accepted patterns are inevitable. In order to reduce false acceptance, the association measure is modified as formula (1)

$$association(x, y) = \begin{cases} 0 & f(x,y) < c \\ I(x,y) & I(x,y) \geq t \\ I(x,y) * z_x(y) & f(x) \leq f(y), \sigma(f(x,*)) \neq 0, I(x,y) * z_x(y) \geq t \\ I(x,y) * z_y(x) & f(x) \geq f(y), \sigma(f(*,y)) \neq 0, I(x,y) * z_y(x) \geq t \\ I(x,y) * \frac{\sqrt[3]{f(x,y)}}{z'_{successor}(x)} & \text{otherwise}, \frac{f(x,y)}{\mu(f(x,*))} \geq 0.9, I(x,y) * \frac{\sqrt[3]{f(x,y)}}{z'_{successor}(x)} \geq t \\ I(x,y) * \frac{\sqrt[3]{f(x,y)}}{z'_{predecessor}(x)} & \text{otherwise}, \frac{f(x,y)}{\mu(f(*,y))} \geq 0.9, I(x,y) * \frac{\sqrt[3]{f(x,y)}}{z'_{predecessor}(x)} \geq t \\ I(x,y) & \text{otherwise} \end{cases} \quad (1)$$

The *I*(*x*,*y*) is the association norm, $z_x(y)$ and $z_y(x)$ are the bigram zscore of *f*(*x*,*y*) among *f*(*x*,*) and *f*(*,*y*) respectively, σ(*f*(*x*,*)) and σ(*f*(*,y)) are standard derivations. $z'_{successor}(x)$ and $z'_{predecessor}(y)$ are the shifted and rescaled zscore of the number of successors and predecessors. They are defined as formula (2).

$$z'_{successor}(x) = \frac{z_{successor}(x) - \min z_{successor} + 1}{\alpha - \min z_{successor} + 1} \quad (2a)$$

$$z'_{predecessor}(y) = \frac{z_{predecessor}(y) - \min z_{predecessor} + 1}{\alpha - \min z_{predecessor} + 1} \quad (2b)$$

$$\min z_{successor} = \frac{1 - \mu(successor)}{\sigma(successor)}, \quad \min z_{predecessor} = \frac{1 - \mu(predecessor)}{\sigma(predecessor)} \quad (2c)$$

Observing formula (1), if a pattern is not fit for the former 4 conditions, which means its association norm is not high enough and bigram zscore is not greater than 1. The low bigram zscore results from *f*(*x*,*y*) being relative low or too near with other *f*(*x*,*) or *f*(*,*y*). To judge whether the pattern is in latter case, we divide *f*(*x*,*y*) by bigram mean value. If the result is greater than or equal to 0.9, we then extract the pattern with high frequency and specific usage of successor or predecessor. Because the range of zscore is from negative, zero to positive, its value is shifted to positive and rescaled to appropriate range.

When the extraction convergence, we get some final patterns, i.e. phrases, and some intermediate patterns, which may be meaningful partial phrases or incomplete patterns just used to extend to complete patterns. For example, "公司(company)", "有限公司 (limited company)" is meaningful partial phrases, while "京東路(King east road)" is meaningless, just used to extend to "南京東路 (Nan King east road)". So the incomplete patterns must be discarded. We measure the function $z_{successor}(*)$ and $z_{predecessor}(*)$ for each pattern to judge the intermediate pattern is complete or not. If one of them is less than α, which means the successor or predecessor of this partial pattern is very specific, then this pattern is impossible to be a phrase boundary and should be viewed as incomplete to be discarded.

3. ROBUST PHRASE SPOTTING

There are many state-of-the-art word spotting techniques, but we present a quite different technique. The presented phrase spotting approach is performed on the syllable lattice, which is recognized based on the acoustic front-end of Mandarin continuous speech recognition. In order to solve the problems of insertion/deletion/substitution from syllable recognition errors and extra modifiers or abbreviations by user, we design a robust identification technique on syllable level for spotting phrases. For each phrase, we generates a "time span graph" recording the beginning and ending time frames each syllable occurred on syllable lattice, then identify phrase boundary by searching the *shortest path* on graph with a Viterbi algorithm [6].

Though the test results prove that the previous work is efficient, observing the spotted phrases, the spotting rate is high but the false alarm rate is not very low. These false spotted phrases will increase the possibility of understanding errors. Here, we propose some improvements. First, every component syllable in "time span graph" for each phrase generated from syllable lattice is added one dummy node, representing a substitution. This gives the chance that choosing dummy node instead of other time frame node on the optimal path. In other words, sometimes a phrase is more likely having one substitution than some insertions. Second, the cost function for shortest path search is promoted to n-gram distance.

$$cost(\mathrm{state}_i^{m,n}) = \min_l \left[cost(\mathrm{state}_{i-1}^{l,m}) + \sum_{j=2}^{i} d(t_b(s_i^n), t_e(s_{i-j}^k), j) \right] + d(t_b(s_i^n), t_e(s_{i-1}^m), 1) \quad (3)$$

The *k* is the *k*-th candidate of (*i*-*j*)-th syllable within a phrase, $d(t_b(s_i^n), t_e(s_{i-j}^k), j)$ denotes the time frame difference between the beginning time frame of s_i^n and the ending time frame of s_{i-j}^k. The summation means that the cost is measured by all distance between any two syllables within a phrase as shown in figure 3. The definition of $state_i^{m,n}$ is described in [6]. When the distance between two syllables is calculated, if one or two of them are dummy nodes, which means there exists deletion or substitution on the searching path and we give them *lost* values to represent distance. For the value of $d(t_b(s_i^n), t_e(s_{i-j}^k), j)$, if it is lower than 1, which means $< s_i^n, s_{i-j}^k >$ are contiguous, otherwise there exists insertion on the path $< s_i^n, s_{i-j}^k >$. Third, we further define the score of phrases by integration of the costs and acoustic scores as shown in formula (4).

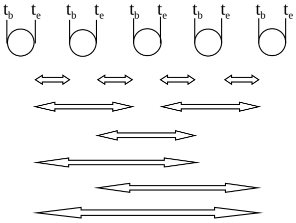

Figure 3: The cost for a syllable path in time span graph matching a phrase.

$$\begin{aligned} PhraseScore &= \lambda * similarity + (1-\lambda) * AcousticScore \\ &= \lambda * \frac{1}{\min\limits_{m,n} \cos t(state_L^{m,n}) + 1} + (1-\lambda) * AcousticScore \end{aligned} \quad (4)$$

Symbol L is the length of phrase (number of syllables). Because the cost is a minimum estimation, its inverse represents the similarity between the spotted segment and the phrase, and it is added one before inverse to avoid dividing by zero. Thus the range of *similarity* is between 0 and 1 and consistent to the range of probability. Besides, the acoustic scores are processed by simple verification, *Sigmoid* function [7]. Those syllable hypotheses with verification score lower than a threshold are rejected. The *Sigmoid* function not only verifies the syllable hypotheses but also appropriately reduces score range to (0,1) to be consistent to the range of probability.

Connecting the phrase hypotheses spotted by robust phrase technique, we can get a phrase lattice. Since the phrase lexicon is the significant and meaningful parts extracted from corpus, there still have insignificant parts in sentence. We view them as fillers. The final phrase lattice with fillers is as the example in figure 4.

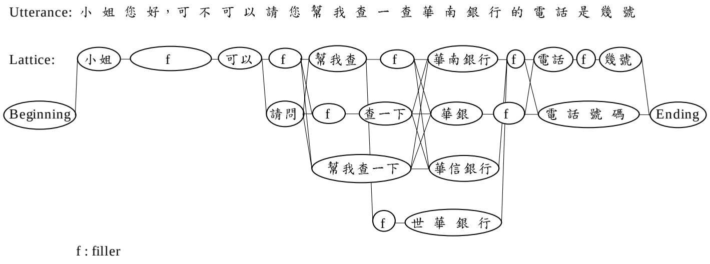

Figure 4: Phrase lattice example.

4. STATISTICS-BASED SEMANTIC PARSING AND INTENTION IDENTIFICATION

During understanding phase, robust phrase spotting is applied on the syllable lattice (result of recognition phase) and then the phrase-level parsing is performed on the phrase lattice to get the complete meaning. Since a sentence is not composed of **many** phrases, the parsing process can be well handled by phrase-level n-gram language models, which integrate phrase, semantic concept and dialogue model to model intra and inter sentence structure. Finally, intention identification is used to classify the complete meaning to a higher-level intention abstraction via probabilistic finite state network for further processed by dialogue manager.

4.1. Phrase Classification

Every phrase has its meaning in semantics. For example, “Please help me find”, “I want to inquire”, or “May I ask” all represent inquiries; “What”, “Pardon”, or “Would you please repeat again” all represent unclarity. We can understand speaker’s intention from the combination of the meanings of all the spotted phrases in an utterance. For a specific task, we can define some semantic tags and label a tag for each phrase. But when the task is very complicated, these works become very laborious and tedious. On the other hand, it can not be ported to other tasks directly. Here we want to automatically classify the phrases to some concepts; that is, the phrases with similar meaning in semantics are grouped into the same concept.

First, create a feature vector for every phrase. For a phrase, p, its feature vector, $V(p) = [f(p_1, p), f(p_2, p), \ldots., f(p_n, p), f(p, p_1), f(p, p_2), \ldots., f(p, p_n)]$, n is the phrase lexicon size, the first n dimension is the frequency of all phrases preceding it, and the last n dimension is the frequency of all phrases succeeding it. Second, use vector quantization (VQ) to cluster the n vectors. That is, those phrases with similar predecessors and successors probably have the same semantic meaning. Here we present a modified algorithm which is *not to preset the number of clusters, and not necessary to be exponential of 2*. The modified algorithm begins with one cluster and **splits one more cluster** every iteration until the *average distance* falls below a threshold. After the modified VQ processed, the phrases with similar meanings are clustered together. We call the clusters as concepts. To realize their meanings, we give each concept a tag name.

4.2. Phrase-Concept Joint Language Model

Because of the speech recognition ambiguities and errors as well as the inherence of robust spotting approach, the false alarms resulted by phrase spotting are unavoidable. These false phrases may lead to misunderstanding. Therefore, it is important to reduce the false alarms for correct understanding. It is well known that conventional word n-gram language model used in the linguistic processing of speech recognition achieves good performance. We follow the point to present a phrase-concept joint bigram language model, which is able to perform syntactic and semantic checking for modeling the intra sentence structure and rejecting the false phrases.

For acoustic observation, O, we search for the top N phrase path $P = (p_1, p_2, \ldots\ldots, p_k)$ on phrase lattice with corresponding concept path $C = (c_1, c_2, \ldots\ldots, c_k)$ and the fillers F.

$$\begin{aligned} \arg\underset{P,C}{topN}\ \Pr(P,C,F \mid O) &\cong \arg\underset{P,C}{topN}\ \Pr(O \mid P,C,F)\Pr(P,C,F) \\ &\cong \arg\underset{P,C}{topN}\ \Pr(O \mid P,C,F)\Pr(P,C) \end{aligned} \quad (5)$$

Since the language model is not related to fillers, $\arg\limits_{P,C}\Pr(P,C,F) = \arg\limits_{P,C}\Pr(P,C)$. $\Pr(O \mid P,C,F)$ represents the combination of *PhraseScore* described in section 3 and the score of fillers. Its log value is defined in formula (6).

$$\log\Pr(O \mid P,C,F) = \sum_i \log PhraseScore(p_i) + \sum_i \sum_{t_e(p_{i-1})+1}^{t_b(p_i)-1} \log\delta \quad (6)$$

$\Pr(P, C)$ is the phrase-concept joint language model formulated below.

$$\begin{aligned}\log \Pr(P, C) &= (1-\alpha) \log \Pr(P) + \alpha \log \Pr(C) \\ &= (1-\alpha)\sum_i \log \Pr(p_i \mid p_{i-1}) + \alpha \sum_i \log \Pr(p_i \mid c_i) \Pr(c_i \mid c_{i-1})\end{aligned} \tag{7}$$

4.3. Concept-Based Dialogue Model

In order to more precisely model the inter-sentence relation in dialogue model, a concept-based dialogue model is developed. The unit of dialogue model is based on the semantic unit, concept, instead of whole sentence as usual. We want to model the semantic relation of inter-sentence instead of the relation of intention abstraction. So every concept in searching phrase lattice is conditioned to the concept sequence of previous two utterances (one is speaker utterance, the other is system response) as shown in formula (8). Integrating the concept-based dialogue model to rescore the top N phrase/concept paths defined in formula (7) and then decides the final top1 path, of which the intention of utterance is composed.

$$\begin{aligned}\log \Pr(P, C) &= \alpha \log \Pr(P) + \beta \log \Pr(C) + \gamma \log \Pr(C \mid C^{-2}, C^{-1}) \\ &= \alpha \sum_i \log \Pr(p_i \mid p_{i-1}) + \beta \sum_i \log \Pr(p_i \mid c_i) \Pr(c_i \mid c_{i-1}) \\ &\quad + \gamma \sum_i \log\left(\sum_{j=1}^{J} \sum_{k=1}^{K} \Pr(c_i \mid c_k^{-2}, c_j^{-1}) \big/ JK \right)\end{aligned} \tag{8}$$

In above formula, c_i, is the corresponding concept of *i*-th phrase on the current searching path, c_j^{-1} is the *j*-th concept on the concept sequence of the last system response, and c_k^{-2} is the *k*-th concept on the concept sequence of the second preceding utterance, i.e. last speaker's utterance. *J* and *K* are the number of phrases in previous two utterances.

4.4. Intention Identification

By the language and dialogue models estimation, we get the maximum likelihood concept path. Now we try to identify the intentions conveyed in the concept sequence by finite state networks (FSNs). The concept sequences with the same intention abstraction in corpus are trained to form a FSN. Each FSN represents a kind of speech act type (SAT). The intention annotation for training corpus allows more than one SAT in a dialogue turn. The FSNs are applied to do higher level pragmatic grammar checking. Because the FSNs are probabilistic, there always have a maximum likely intention even out of grammar. Besides, if more than one FSN accepted in checking, which means the speaker expresses more than one speech act. So the final multi-intention speech is composed of some SATs, which will be further processed by the dialogue manager and associated with database management.

5. EXPERIMENTS

A recently completed successfully working Chinese spoken dialogue system for telephone directory services is developed. The task is for all banking/financing organizations in Taipei, with a total of 4208 phone numbers. This system is primarily trained with a corpus recorded from real conversations between human beings, therefore is user/system mixed-initiative simulating the two-way dialogue to a good extent about 21 kind of intentions, including greetings, inquiring telephone numbers, requesting divisions/ departments or other lines, asking for idle telephone numbers, repeating numbers or asking for repeating utterances, etc. The system accepts either a syllable recognizer or a Chinese keyword spotter as the acoustic front-end. Most of the models are trained from the corpus with bootstrapping strategy, thus are flexible with good portability to different tasks when the corpus is available. In initial experiments the training corpus includes transcriptions of 1156 human dialogues with a total of 12,776 sentences and 88,119 characters. They were obtained from Chung-Hua Telecom in Taiwan recorded from real human-to-human directory services. In the tests, 77.99% of top15 candidate inclusion rate for syllable recognition front-end gives 79.39% of phrase spotting rate and 80.42% of user intention estimation accuracy. Ignored the out of task utterances, the accuracy can achieve 87.89%. Further improvements for the system are currently under progress.

6. Conclusions

We successfully developed a telephone directory service spoken language dialogue system for Mandarin Chinese. It is a user/system mixed-initiative dialogue system to simulate the conversation may occur in client-agent telephone directory services to a good extent. The proposed statistics-based approach is capable of modeling speaker's intention and integrating human knowledge. The test results prove that the proposed approach is efficient and can be easily applied to various spoken dialogue applications.

ACKNOWLEDGMENT

The authors acknowledge the provider of Mandarin speech acoustic recognition front-end by Chen, Berlin at Institute of Information Science, Academia Sinica, Taiwan, ROC.

REFERENCES

[1] Wang, C., Zue, V., *et al*. "YINHE: A Mandarin Chinese Version of the GALAXY System," *Proceedings of EuroSpeech*, pp. 351-354, Rhodes, Greece, 1997.

[2] Zue, V. "Conversational Interfaces: Advances And Challenges," Proceedings of EuroSpeech, pp. KN9-KN18, Rhodes, Greece, 1997.

[3] Seneff, S. "Robust Parsing for Spoken Language Systems," Proceedings of IEEE ASSP, pp. 189-192, Minneapolis, USA, 1993.

[4] Briscoe, E. J. and Carroll, J. "Generalized Probabilistic LR Parsing of Natural Language (Corpora) with Unification-Based Grammars," Computational Linguistics, Vol. 19, No. 1, pp. 25-59, 1993.

[5] Yang, Y. J., Chien, L. F., and Lee, L. S. "A New Efficient Approach for Collocation Extraction," Proceedings of ICCPOL, pp. 484-487, Hong Kong, 1997.

[6] Yang, Y. J., Chien, L. F., and Lee, L. S. "Speaker Intention Modeling for Large Vocabulary Mandarin Spoken Dialogues," Proceedings of ICSLP, Vol. 2, pp. 713-716, Philadelphia, USA, 1996.

[7] Lleida, E., Rose, R. C. "Likelihood Ratio Decoding and Confidence Measure for Continuous Speech Recognition," Proceedings of ICSLP, Vol. 1, pp. 478-481, Philadelphia, USA, 1996.