International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages, pp. 225-228, Mar. 28-30, 2004, Beijing, China

Context Directed Speech Recognition in Dialogue Systems

Pengju YAN[1] *and Fang ZHENG* [2]

Center for Speech Technology, State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science & Technology, Tsinghua University, Beijing, 100084, China
{yan, fzheng}@cst.cs.tsinghua.edu.cn, http://cst.cs.tsinghua.edu.cn

Abstract

For the time being, research and development on spoken dialogue systems (SDSs) become more and more important with the constant increasing demands. However, defects of the recognition strategies adopted for the sake of only laboratory demos are revealed evidently under the real-world circumstances, with the utterances of the casual style instead of the declamatory one. Both the shortage of domain-specific corpus and the existence of other empirical/heuristic knowledge appeal new methods to improve the recognition performance. Here we present a recognition framework where the dialogue contexts (DCs) are incorporated in as a restrictive source. Firstly, the idea of a focus expected (FE) under certain dialogue states is introduced. Secondly, the adaptation of lexicon and grammar rules is proposed. Finally, the recognition automaton generation under a specific FE is put forward. Experiments are carried out in the dialogue system *EasyFlight*, and the results show the effectiveness of the strategies.

1. Introduction

In a somewhat inaccurate way, the term *spoken dialogue system* (or *dialogue system* in brief) can be defined as an automatic service providing system via the speech interaction I/O interface to people. Just as that implies, a dialogue system normally consists of four functional components which are a speech recognizer, a language parser, a dialogue manager and a speech synthesizer. Differing from in-laboratory speech systems, the main goal of a dialogue system is to achieve a real-world pragmatic task, e.g. to find out a best route to a site, or to book an air ticket. Thus the understanding performance becomes the most cared issue that researchers focus on.

Measures have been taken to counter the spontaneousness/ casualness of the spoken utterances in dialogue systems. At the pure acoustic level, hybrid pronunciation modeling is proposed to deal with rich pronunciation variants and notable co-articulations, and primary but encouraging progress has been made, such as given in [1]. At the pure linguistic level, a robust understanding scheme which models repeat, word disordering, fragment, ellipse, and ill form is present in [2], and the grammar coverage is proved to be sufficiently large against all of those ungrammaticalities. However, the recognition strategy itself still plays a bottleneck role in the whole scene.

Generally speaking, there are four kinds of speech recognition strategies appearing by now that can be used in dialogue systems. The first and the simplest way is the isolated word recognition. Owing to the high recognition rate, it can be used in crucial situations where even the least errors are not tolerable, but shows very low user-friendliness. The second is the keyword spotting, where the main idea is to highlight the task-concerned words comparing with the unconcerned ones by means of using various weights [3]. One hybrid is the so-called *sliding-window* word spotting, where the search process can start at anywhere in the speech [4]. The main disadvantage of them is that any other knowledge can only be adopted as a confidence-measure, thus produces low recognition rate. The third is the template based matching where the input utterances are explicitly restricted in the search graph [4]. Even if the network is expanded at arcs by altering words within the same semantic class, the performance is rather low against unpredicted utterances. The fourth is the stochastic n-gram based recognition. Considering the short of sufficient training corpus, unified language models integrating n-gram and grammar rules have been put forward [5]. With the perplexity considerably dropped, unfortunately, the word error rate stays almost unchanged.

In this paper, a context directed recognition strategy is proposed, where the dialogue context knowledge and semantic knowledge, what the previous methods neglect, are made best of to instruct the search process. The main idea is to predict the information the next turn will be involved in and then restrict the search process to the predetermined word network. The strategy can be depicted as follows. At first, a *focus expected* is introduced in each dialogue turn to reflect the current dialogue inner status, which is the function of the *history/context*. Next, given a specific FE, a rule set is dynamically chosen according to the offline semantic label. Once the rule set related to each FE satisfies some condition, it can be converted to a *finite state network* (FSN) with its arcs associated with words. Finally, the recognizer produces the ultimate results by searching through the given FSN. This strategy is tried on an air travel information

[1] The author is now with Panasonic Beijing Laboratory. mailto:\\yanpj@cmrd.panasonic.com.cn

[2] The author is also with Beijing d-Ear Technologies Co., Ltd. mailto:\\fzheng@d-Ear.com

query and booking system of *EasyFlight* and the satisfying results are achieved.

The rest of this paper is organized as follows. The idea of the context directed recognition framework with the definition of a FE is firstly introduced in Section 2. The construction of a FSN under a specific FE as well as its usage in the recognition stage is described in detail in Section 3. Experimental results are presented in Section 4. And finally, conclusions are drawn in Section 5.

2. Context directed recognition

The prevalent speech and language processing frameworks divide the speech understanding task as the combination of some isolated models/modules, which are acoustic model, language model, grammar, semantics, and dialogue management, and data are passed and processed in strict sequence from begin to end. However, such knowledge in different levels has inherent relations to each other. The integration of (some of) them into one stage is expected to be beneficial, such as language model look-ahead technique helps to raises the overall performance. Consider that in real world all relative knowledge, such as dialect/accent, syntax/semantics, discourse context, and education back ground, can help people to understand a sentence, the use of them in automatic speech recognition is naturally promising. The solution here mainly involves dialogue context and semantics while leaving other knowledge not discussed.

2.1. Focus expected and its evolvement

In a mixed-initiative dialogue system, it is naturally to assume that the user is cooperative in order to accomplish a real task. In other words, in each userís turn, the user is most probably about to tell the information absent till now or answer the systemís question faithfully. Hence the prediction of next turnís information becomes conceivable, and the mechanism is called a *focus expected*.

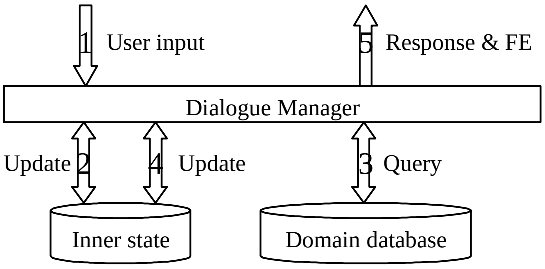

Figure 1: *Focus expected evolvement* [2]

A plan-based dialogue management structure, named topic forest, is adopted in *EasyFlight* [6]. The structure provides a unified representation of multi-topic issue, topic changing issue, information sharing across topics, and the different importance of items. The focus expected is designed to be the function of the dialogue *history/context* at each turn, where the *history/context* can be defined as which information items, associated with their occurring sequence, has been interactively talked about currently. In this way, a *focus expected* is determined at each dialogue turn according to the context to predict what the user will speak next.

A FE is changing along with the dialogue advancement, which is called the FE *evolvement*. There are four kinds of FE evolvements exist in the dialogue manager. The first is that when some necessary information is absent, the system will ask about that; the second that when certain information has more than one optional values, the system will ask the user to choose among them; the third that when all information is determined, the system ask the user to confirm them one by one; and the last that when history stays unchanged, the FE stays unchanged too. The process of a FE evolvement is depicted in Figure 1, where the numbers inside the hollow arrows stands for the order of the data flow.

2.2. Framework

Given the focus expected as context information, the semantic prediction of the next sentence becomes feasible. It is the language understanderís turn to do that especially when it adopts a rule-based parsing technique. To be robust against various ungrammatical linguistic phenomena in spontaneous speech, a hybrid rule-based understanding scheme is proposed by us in [2,7], where five types of rules are introduced to describe the ungrammaticalities. In *EasyFlight*, keyword and semantic categories are used as grammar symbols. An enhanced chart parser, *marionette*, is designed to implement the advantage of the extended grammar. It is approved that the strategies enlarge the grammar coverage. Moreover, it provides as well a straightforward way to perform the semantic prediction.

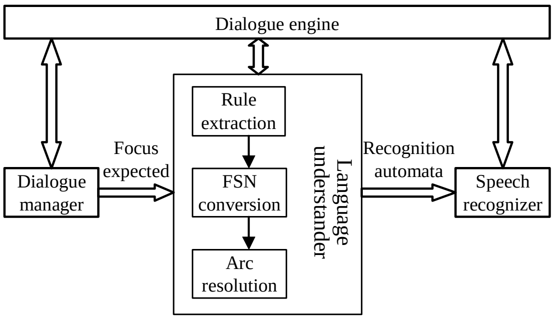

Figure 2: *Framework of context directed speech recognition* [2]

In Figure 2, the framework of the context directed speech recognition is depicted, where the language understander plays an important role by carrying out rule set and lexicon adaptation. The detail will be given in the next section.

3. FSN recognition automata

The focuses expected can be classified according to their semantic differences. And as mentioned above, the system grammar is transcribed by using keyword and semantic categories, thus forming a semantic grammar. Therefore, the semantic phrasal patterns under different FEs can be modeled by different subsets of rules of the grammar. In the domain of *EasyFlight*, the semantic classes comprise: *location*, *flight no.*, *date*, *time*, *airplane type*, *airways name*, *ticket account*, *personal ID no.*, and *confirm phrase*. The focuses expected can be one of or the combination of those classes. Semantic classes are labeled onto each rule, and all the rules belonging to a specific FE composes an active rule set. The active rule set can be converted to be a FSN if some criteria met, and the grammar symbol attached to arcs can be further resolved to be words/fillers.

3.1. Semantic class labeling

The semantic classes are assigned to each rule manually. The labeling follows several criteria:

✧ The generative rules within one semantic class must be equivalent to a FSN in terms of description power, i.e., not include recursive rules;

✧ To keep the automaton concise and easy to handle, exclude the rule types of *long-spanning*, *up-messing*, and *over-crossing*, where the modeled linguistic phenomena are too complex [2]; and

✧ The rule sets of a semantic class can only describe phrase level structures rather than sentence level structures, to avoid a too strong search restriction.

dgt_h → ato_0	[*4, 5*]
dgt_h → ato_1_10_l	[*4, 5*]
*dgt_h *→ ato_10 ato_1_9*	[*4, 5*]
*dgt_h *→ ato_2 ato_10*	[*4, 5*]
*dgt_h *→ ato_2 ato_10 ato_1_3*	[*4, 5*]
sub_from → mat_city_name	[*0, 1*]
sub_from → tag_from_here	[*0, 1*]
sub_from → tag_from mat_city_name	[*0, 1*]
sub_stop → tag_stop mat_city_name	[*0, 1*]

Figure 3: *A labeling fraction*

An extracted labeling example is listed in Figure 3, where 0/1 stand for the departure/arrival city classes and 4/5 the departure/arrival time classes. The production symbol with its optional proceeding character indicates the type of the rule. ì *→î indicates the ordinary rule type, called *up-tying*, as in a normal CFG; while ì →î indicates the *by-passing* rule type, by which sub-constituents can be grouped together by skipping a number of segments [2,7]. Note that terminal symbols are also the keyword categories in the lexicon.

3.2. FSN building

The FSN is adopted as the recognition network because of the two advantages: one is that it is convenient to convert from a set of non-recursive rules to a FSN, and the other that the FSN gives a highly predictive way to restrict the search space by leaving out out-of-rule hypotheses.

*A *→ aa*
B→ Abc
*C *→ cc*

Figure 4: *An active rule set example* [2]

Given the focus expected, an active rule set is extracted from the grammar by matching the semantics of the FE and each rule, remember that only non-recursive and *up-tying/by-passing* rules included. To convert that rule set to a FSN, standard conversion processes must be modified to deal with the *by-passing* rules. For a *by-passing* rule of *B → Abc* in Figure 4 the corresponding network for *B* is given in Figure 5, comparing with the conversion of a up-tying rule of *C *→ cc*. Once all active rules are processed, the network is integrated to be a non-deterministic finite state automaton (NFA) at the bottom of Figure 5.

Besides lexical items, such as *a*, *b*, and *c*, the special sign items of ε, ϕ, and ω attached to the arcs need to be further explained. The item ε denotes a null arc and will be eliminated afterwards. The item ϕ denotes a filler arc where a special set of filler words, such as ì*en*î (hesitate voice in Chinese somewhat similar to *um* in English), can get through. And the item ω denotes an inactive arc where all FE-unconcerned words (to be defined in the next subsection) can propagate. In contrast to a template based matching method, our recognition network here allows more freely uttered and unconcerned phrases to be pass through the network. Thus keywords not matched with the FE can also been recognized, which facilitates the mixed-initiative dialogue strategy.

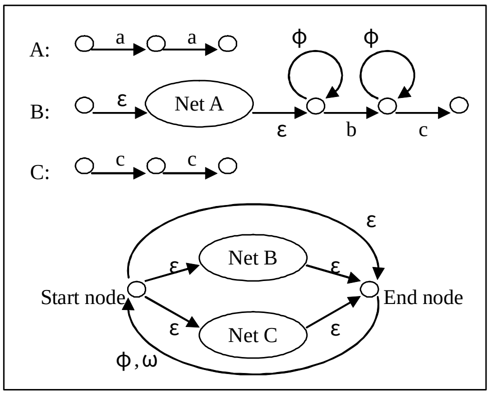

Figure 5: *Non-deterministic finite automaton conversion* [2]

To remove redundancies and to improve efficiencies, the NFA is subsequently transformed to a determined finite state automaton (DFA), referred to a determined FSN in this paper, and minimized in size.

3.3. Arc resolution

The previous steps build the FSN under a given FE, but the symbols attached to arcs are still semantic category items. To give an explicit propagation process, the arc symbols must be resolve to words. In other words, the recognizer should know which words can propagate from one node to another. There three kind of situations. First, a terminal symbol corresponds to the keywords within the keyword category. Words of all the terminal symbols compose the active word set, also named FE-concerned word set. Second, the item ϕ corresponds to a set of filler words. At last, the item ω corresponds to all the rest words except the active and the filler ones making up of the inactive word set, or FE-unconcerned word set. By this way, each symbol clinging to the arcs is replaced by a set of words.

3.4. Use in recognition

In addition to the fact that linguistic knowledge sources are embodied in the FSN, heuristic and stochastic knowledge can also been incorporated in. For example, probabilities can be added to arcs to indicate the reliability of each rule. And phrase level n-grams can be considered when the expansion reoccurring at the start node. Even the automaton has the big advantages, stochastic knowledge is not used in the lack of domain-specific corpus. A weighting heuristic strategy is adopted to highlight the

FE-concerned words/phrases with fillers and inactive words in the background, i.e., FE-concerned words, FE-unconcerned words, and fillers are given different pre-tuned weights respectively.

The recognition automaton under a FE is customized at dialogue turn and the output word lists is transferred to the language understander to compute the corresponding semantics.

4. Experimental results

The experiments are made in the domain of *EasyFlight*. A set of 42-dimensional MFCC-based features are adopted, and the acoustic models are trained from the standard 863 assessment, where the 520 sentences for each of the 70 speakers are uttered in a declamatory manner. The test data contains 100 sentences for each of the 5 speakers and is of the spontaneous/casual style. Both the training set and the test set are 16k Hz sampled through PCs under laboratory environments.

4.1. Design of experiments

Four representative focuses expected are included in the experiments. A). A null FE FE_{null}, which is not a realistic FE in *EasyFlight*, degrading the recognizer to be a keyword spotter. B). A time FE FE_{time} integrating 90 rules describing date/time expressions, such as ì *X月Y日(号) ñ X month Y day*î and ì*X点 Y 分 ñ X hour Y minute*î. C). A time-location FE FE_{time_loc} composed of FE_{time} and 13 more rules describing location expressions, such as ì*从X到 Y ñ from X to Y*î. And D). an ordinary FE FE_{full}, also a non-realistic FE in real systems, putting together mid-granular rules describing all task-concerned concepts.

Semantic units in the test corpus are marked out in order to assess the recognition rate of FE-concerned units under each FE. Words of the corpus are then divided into out-of-vocabulary words *O*(9.1%), fillers *F*(8.4%), interrogatives *B*(6.9%), time units *T*(20.8%), location units *L*(14.5%), and other units *I*(40.3%), where each percentages in the parentheses indicates the length in syllable. The relationship between FEs and semantic units are FE_{null} = B+T+L+I+F, FE_{time} = B+T, FE_{time_loc} = B+T+L, and FE_{full} = B+T+L+I.

4.2. Results and analyses

Four measures are considered in the experiments, which are the syllable correction rate (SCR), syllable accuracy rate (SAR), FE-concerned SCR (FECR), and FE-concerned SAR (FEAR).

Similar to a keyword spotting strategy with keywords and fillers paralleling but differently weighted in the network, the SCR, SAR, FECR, and FEAR under a FE_{null} FSN are 70.8%, 69.2%, 75.6%, and 73.0% respectively. It is an acceptable approximation that different semantic units share the same recognition rate the whole sentence achieves, therefore we take them as the baselines when discuss the performances under the three other FEs.

In Table 1, the error reduction rates (ERRs) of the four kinds of measures are listed. We achieve syllable ERRs by approximately 10% as show in the 2nd and 3rd row. It means that the overall performances are considerably improved when semantic connections are restricted in the recognizer. We also achieve great FECR ERRs by about 35% when considering non-ordinary FEs, as shown in the 4th row. While the FEAR ERRs are not satisfying, the language understander can deal with the insert errors very well, thus alleviates the infection of that.

Table 1: *Experimental results*

ERR(%)	FE_{time}	$FE_{time\ lo}$	FE_{full}
SCR	11.1	16.4	7.5
SAR	10.9	15.4	7.4
FECR	34.0	38.9	-14.3
FEAR	-22.6	5.1	-41.9

We can see that FEAR and FECR ERRs are bad under the virtual FE of FE_{full}, that is because the too complex active rule set produces a too perplex FSN. It is also the reason why the FECR under FE_{time} is a minus value, considering the corresponding rule set size of which is 90. However, when a small rule set is added to FE_{time} to make FE_{time_loc}, the FECR greatly rises because of the perplexity reduction between the two FSNs. It can be learned from the observations that the FE should better be comparatively explicit to make a concise FSN.

5. Conclusions

In this paper, a dialogue context directed recognition strategy is proposed. It aims at using context and semantic knowledge sources to restrict the recognition search space, in the face of the common difficulties to get sufficient domain-specific corpus for a stochastic approach. While recognition rates of FE-concerned semantic units under dialogue context are greatly improved, it also provides a flexible way to let FE-unconcerned words to get through, which facilitates the mixed-initiative dialogue strategy.

6. References

[1] Song, Z.-J., 2001. Research on Pronunciation Modeling for Spontaneous Chinese Speech Recognition. *Ph.D. Dissertation*. Tsinghua University, P.R.C.

[2] Yan, P.-J., 2002. Research on Natural Language Understanding in Dialogue Systems. *Ph.D. Dissertation*. Tsinghua University, P.R.C.

[3] Zheng, F., 1997. Studies on Approaches to Keyword Spotting in Unconstrained Continuous Speech. *Ph.D. Dissertation*. Tsinghua University, P.R.C.

[4] Lu, Z.-Z., 2002. Research on Speech Recognition for Spoken Dialog system. *M.Eng. Thesis*. Tsinghua University, P.R.C.

[5] Wang, Y.-Y.; Mahajan, M.; Huang, X., 2000. A Unified Context-Free Grammar and N-Gram Model for Spoken Language Processing, *Proceedings of ICASSP2000*. Istanbul, Turkey.

[6] Wu, X.-J.; Zheng, F.; Xu, M.-X., 2001. Topic Forest: A Plan-Based Dialog Management Structure. *Proceedings of ICASSP2001*. Salt Lake City.

[7] Yan, P.-J.; Zheng F.; Xu, M.-X., 2001. Robust Parsing in Spoken Dialogue Systems. *7th European Conference on Speech Communication and Technology*. 2149-2153, Aalborg, Denmark.