ALICE : ACQUISITION OF LANGUAGE IN CONVERSATIONAL ENVIRONMENT

— An Approach to Weakly Supervised Training of Spoken Language System for Language Porting —

Tetsunori Kobayashi

EECE, Waseda University

ABSTRACT

Aiming at reducing the work required for the language porting of spoken language system, a conversational second language acquisition system is proposed. This system need only small lexicon in the initial stage. It need neither hand-description of rules nor the collection/annotation of large corpus. It refer the corpus of semantic frames which is obtained through development/use of first language version of the system. Then, it make hypotheses which leed to reasonable semantic frames and parse the sentence with them. The system drive the back-end system with the interpretation and confirm if the result is suit for the user's will. With above process, the weakly supervised training of the spoken language system is realized.

1 INTRODUCTION

1.1 ALICE Assist Language Porting with Weakly Supervised Training

The current version of ALICE works as a second language acquisition system, namely we assume the ability of the system to understand another language. The main propose of ALICE is to reduce the work which is required for porting the existing spoken language system to another language. Especially, it is aiming at reducing the work for the setup of the linguistic knowledge by automatic acquisition through interactive conversation.

The system does not require any training corpus nor pregiven rules for the target language. The system use only the corpus of English (non-target language) which have been used for developing the first language system. The only information for supervising the system is the answer for the YES/NO question asking if the result is correct or not. If the response is positive, the system recognize the hypotheses used for getting the interpretation are correct and register them to its knowledge-base. The system gain its knowledge under such a weakly supervised condition.

1.2 Why Weakly Supervised Training

Recently, many spoken language systems have been developed. However, they are task and language dependent without exception. If we want to use it in different condition, say different task or different language, we have to take long term to rebuild the system through tremendous training procedures. Training is always the bottleneck for building a new system. One reason why the training is so hard is the current training procedure need "strong supervising". To reduce the un-neglectable work, it is desired for the training procedure to be more "weak". Namely, we need training procedure which scarcely bother the system builder.

1.3 Why Linguistic Knowledge Acquisition

ALICE focus on the acquisition of linguistic knowledge (here, I use the word of "linguistic knowledge" for the meaning of the grammatical and lexical rule used for getting semantic information from given word sequence).

Conventionally, we have tried to get linguistic knowledge by rule-based approach or corpus-based approach. In the rule-based approach, we write new rules by hand. In the corpus-based approach, we collect large amount of data and annotate them. First approach requires the expert to the write rules. and second one requires the tedious works. So, the automatic linguistic knowledge acquisition should be very attractive.

There are many other resources to be churned when we port the spoken language system: acoustic model and language model. As for the acoustic model, there are many speech corpus for many language these days. Therefore, this part has little problem for the porting. As for the language model, there is some possibility to get it from linguistic knowledge. So the linguistic knowledge acquisition should be very essential for the language porting problem.

1.4 What's new in ALICE

First novelty of ALICE is the reuse of the text corpus of non-target language. Since the goal of our system is language porting, we can expect the text corpus of the first language which have been used for the development of the first system. This corpus can be easily translated to the corpus of semantic frames. Our system try to parse the input sentences with only this corpus, while almost of all other systems use the data consist of the text in target language and its semantic representation.

The parser which enables this feature (namely the parser which can induce the grammatical rules and word meaning

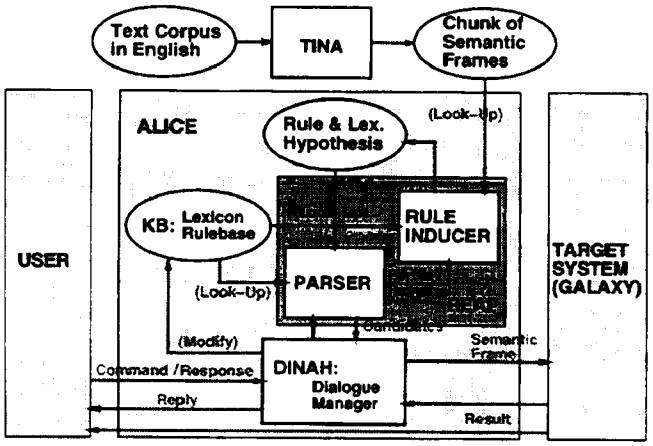

Figure 1: System organization of ALICE.

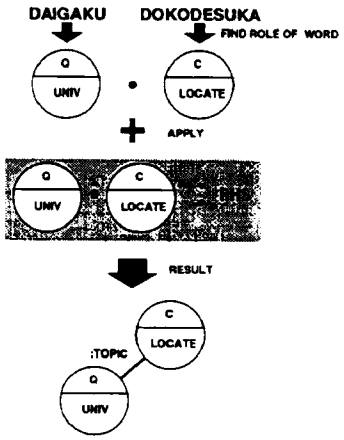

Figure 2: Example of rule application (RHS).

with only semantic corpus) is the second novelty of ALICE. This function is performed by estimating the rule which can produce a semantic frame similar to those which appear in the corpus of semantic frame.

Many alternatives may be obtained in this framework. In this case, supervisor teaches the system which interpretation is correct. Here, I propose a new supervising scheme, that is the supervising through execution. This is the third novelty.

Through these processes, the weak supervising is realized.

2 THE ALICE SYSTEM

2.1 The Organization of ALICE

ALICE works as a front-end of the GALAXY system which is a multi-domain Q/A system developed at MIT-LCS[1]. Users can use GALAXY in their mother language through ALICE.

Figure 1 shows the configuration of the ALICE system.

The system consist of two parts. One is the dialogue manager named DINAH (DIalogue Navigator as Alice's Hub), which decide what kind of dialogue mode is suit the situation. The other is the parser named REAP (Rule Estimation Applicable Parser), which can induce the linguistic knowledge such as word meanings and the phrase structure rules. The parser consist of two parts. One is the regular parser based on bottom up chart parsing. The other is the rule inducer which make hypothesis of rules/ lexicon.

This system use only one external database, that is semantic frame database. I call this database case-base. All other database is internal and self-expanding. The case-base can be easily obtained by applying the English text-base which is used to develop the English version of target system to English parser named TINA[2].

2.2 How ALICE Works

DINAH receive the command utterance from user and send it to REAP. Firstly, REAP try to parse the sentence using only current Knowledge-base. if the parsing is successfully finished, then the obtained semantic frame is send to the back-end system through DINAH.

If the parsing is failed, then ALICE move to next phase called EEC process. This process consist of three phases : Estimate, Execute & Confirm by user's response.

In the estimation phase, REAP try to estimate rules/ lexicon which make it possible to parse the sentence using case-base and current knowledge-base(KB).

In the execution phase, REAP try to parse the sentence with KB and estimated rules/lexicon. If the parsing is finished, output candidates are send to the DINAH. DINAH evaluate the likelihood of each candidates and send the candidate with highest likelihood to the back-end system GALAXY. The likelihood is defined by checking the consistency of the constraint laying in the substructure of the frame. Then, the output from the back-end is sent back to the user.

In the confirm phase, DINAH receive the response from the user. If the response is negative, DINAH issue second candidate to back-end. If the response is positive, DINAH add the rules/lexicon which is used to get the accepted interpretation.

2.3 Basic Framework of REAP

REAP is a bottom-up parser with context free grammar. REAP try to generate parse-tree which is likely to appear in the case-base even if the parser cannot find enough rules to parse the sentence,

For the purpose, the rules should be easily maintenanced. Here, the simple rule set using raw word information is adopted. The form of the rule is as follows :

(left word , right word , combine rule, <extra info.>)

Here, 'left word' means 'head word of left phrase'. There are 9 combine rules: RHS(right head standard), LHS, RHG(right head with glue), LHG, IGR(ignore right), IGL, GNT, GFR(generate & fill right), GFL.

Figure 2 shows the example of application of rule RHS.

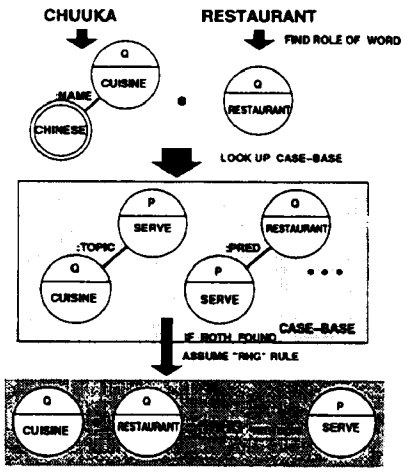

Figure 3: Example of rule estimation (RHG).

2.4 How to Induce Grammatical Rule

2.4.1 Rule estimation using case-base of semantic frames

When REAP fail the parsing with normal rules, it try to make rule hypotheses through the analysis of case-base and current KB. The rule assuming procedure is as follows.

For the sequence of A - B, REAP try to find the sub-structure of <A,B> or <B,A> in the case-base. Here, the notation of <A,B> denotes the phrase structure whose head is A and complement is B. If <A,B> is found, the rule (A,B,LHS) is assumed. If <B,A> is found, the rule (A,B,RHS) is assumed.

REAP also assumes two other rules using the combination of <A, C> and <C, B>. If these sub-structures are found in case-base, then the rule (A,B,LHG,C) is assumed. This rule make the structure, <A, <C, B>>. Similarly, if <B, C> and <C, A> are found, then the rule (A,B,RHG,C) is assumed. This rule make the structure, <B, <C, A>>. These rules are used when B cannot be attached to A directly but B can attached to A through C. The structure C is called "glue" for A and B. Figure 3 shows an example of the rule estimation of this type. In this case, it is found that the node ":pred {p serve }" can be the head of " : topic {q cuisine}" and also can be the complement of ":topic {q restaurant }". Therefore, RHG rule with glue ":pred {p serve}" is assumed.

Using these rules, rule set is expanded and they are applied to the sentence again. If the sentences are successfully parsed, then the assumed rules which is used to generate the correct parse tree of the sentence is registered in the rule-base as the regular rules.

2.4.2 Rule estimation by rule relaxation

There is another way to get rule hypotheses. That is a rule relaxation method. If the system fail to find rule which suit the parsing situation, it try to find similar rule by relaxing the firing condition of rules. Then the rule is modified to be suit for the parsing situation and apply to the situation. If the rule contribute for the parsing of the sentence, then it is registered in the rule-base as the regular rule.

Each method compensate for the fault of the other. The rule relaxation method is rather effective because the rule is borrowed from the existing rules. If the similar rule is exist, the expansion of the rule is easy and this expansion is also applicable to the very new phrase which is not appeared in the case-base. However, at the beginning stage where the rules are not sufficient, this process does not work well because the system tend to fail to find the similar rule in the rule-base. While, the approach based on the case-base of semantic frames is deal with only the relation of phrases which have already appeared in the case-base but it works even if there are no regular rules.

2.5 How to Induce Lexical Knowledge

2.5.1 Word-meaning estimation using case-base of semantic frames

The above section described how the system work to get the knowledge for combining phrases and making new phrases. These knowledge assume that the role of each word is known. This assumption is easy to fail because many different expressions can be used for even one meaning. New word is always used in the system.

In the case of sequence A - *ukw* - B , where A and B are phrases and *ukw* is a unknown word, the system try to find the possible roles of the *ukw* using the combination of the binomial expressions obtained through the analysis of case-base, which is the same information used in the case-based approach for the grammar acquisition. The possible roles are defined as the common set of the set of A's heads and the set of B's complements, and the common set of the set of A's complements and the set of B's heads. In former case, the head of the total phrase is B and the later case the head is A.

If the parsing is successfully finished, the word hypothesis which lead to the solution is adopted as the meaning of the word.

Figure 4 shows an example of the word hypothesis of this type. In this case, it is found that the node ":toplevel {c locate }" can be the head of " : topic {q university }" and also can be the complement of ":sentence {s sentence }". Therefore, ":toplevel {c locate }" is assumed to be the meaning of word DOKODEUKA.

2.5.2 REAP deal with different meaning of known word

When some meanings of a word are given from the lexicon, the parser give the priority to these meanings. It does not try to estimate the other meanings of the word. This is the first phase analysis. In this phase, the word-meaning-estimation process is applied to only unknown words which is not appeared in the lexicon. However, first phase sometimes fails because some words have another meaning which is different from the meaning in the lexicon. If the first phase is failed, then the system try to parse the sentence with the estimation of another meaning of each suspicious word. Using this two pass method, system can deal with multi meanings of the word.

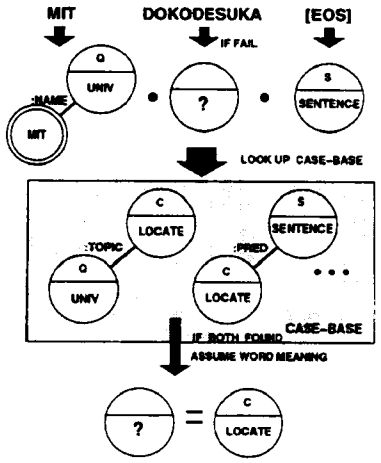

Figure 4: Example of word-meaning estimation.

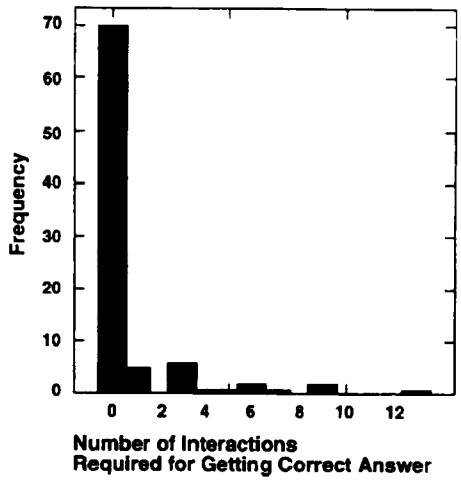

Figure 5: Histogram of the number of interactions required for getting correct answer.

Table 1: Initial/Final condition of knowledge base

Item		Initial	Final	difference
Rule-base		0	146	146
Lexicon	proper noun	75	75	0
	noun	63	63	0
	verb	15	22	7
	functional word	21	62	41
	total	174	222	48
Binomial Expressions		1186	1207	21

3 EXPERIMENT

3.1 Experimental Setup

The 100 English sentences are randomly selected from English corpus which is used for the development of the English version of GALAXY, and then, they are translated into Japanese by hand. They are sorted in small order using the number of words in each sentence. Then the sentences are applied to the ALICE system.

3.2 Experimental Results

As the result, 70 sentences are correctly parsed without interaction, and 19 sentences are correctly parsed after several interactions. As for 11 sentences, parser cannot produce their correct semantic frame.

Histogram of the number of interaction required for getting correct answer is shown in Figure 5.

The initial condition of Knowledge-base and final one after the parsing of 100 sentences is described in Table 1.

4 CONCLUSION

In this paper, I proposed a language acquisition system ALICE. ALICE can parse sentences with insufficient knowledge. Only the case-base of semantic frames and a small basic lexicon are used to estimate linguistic knowledge. All other case-based approaches require the pair of surface representation of sentence and deep structure of the meaning. The characteristics of ALICE that no surface representation is required can be regard as very unique feature.

The evaluation test is performed for randomly selected 100 sentences. As the result, ALICE generate correct semantic frame for 89 test sentences.

Thus the language acquisition system which support weakly supervised language porting is realized.

ACKNOWLEDGEMENT

This work have been done during my sabbatical stay at Spoken Language Systems Group, MIT. Author would express special thanks to Dr. Zue and other SLSers for their support.

REFERENCES

1) D.Goddeau, E.Brill, J.Glass, C.Pao, M.Phillips, J.Polifroni, S.Seneff and V.Zue, "GALAXY: A Human-Language Interface to On-Line Travel Information", Prof. Int. Conf. on Spoken Language Processing 1994, pp.707-711 (1994).

2) S.Seneff, "TINA: A Natural Language System for Spoken Language Applications", Computational Linguistics, Vol. 18, No. 1, pp. 61-86 (1992).

3) A.L.Gorin, L.G.Miller and S.E.Levinson, "Some Experiments in Spoken Language Acquisition", Proc. Int. Conf. on Acoustics, Speech and Signal Processing 1993, pp.I505-I508 (1993).