PREDICTING NOUN-PHRASE SURFACE FORMS USING CONTEXTUAL INFORMATION

Takayuki YAMAOKA, Hitoshi IIDA, and Hidekazu ARITA†
ATR Interpreting Telephony Research Laboratories, Souraku-gun, Kyoto, JAPAN
†Mitsubishi Electric Corporation, Amagasaki, Hyogo, JAPAN

Abstract

We propose a context-sensitive method to predict noun-phrases in the next utterance of a telephone inquiry dialogue. First, information about the utterance type and the discourse entities of the next utterance is grasped using a dialogue interpretation model. Second, a domain-dependent knowledge base for noun-phrase usage is developed, focusing on the dialogue situations in context. Finally, we propose a strategy to make a set of the appropriate expressions in the next utterance, using the information and the knowledge base. This set of expressions is used to select the correct candidate from the speech recognition output. This paper examines some of the processes creating sets of polite expressions, deictić expressions, and compound noun phrases, which are common in telephone inquiry dialogue.

1 Introduction

A high-quality spoken-language processing system must use knowledge of *dialogue* and *spoken-language*. Using *dialogue* knowledge facilitates understanding and predicting utterances in context. Using *spoken-language* knowledge, that is knowledge about how the speaker expresses what he/she wants to say, makes it possible for the system to recognize and generate the more complex expressions that are normally used in our daily dialogues.

To make language processing in the whole spoken-language processing system more efficient, it is vital how to select the correct speech recognition output in the speech-language interface. The use of discourse-level knowledge is an effective way to do this[6][11]. For example, MINDS[6] applied dialogue-level knowledge, particularly for propositional contents, to predict the expected utterance form for the speech recognition. However, although MINDS showed good results, several problems remain before it can be made into a complete spoken-language processing system:

1. how to construct the dialogue structure for the given domain,
2. how to treat predictive concepts regarding not only the propositional contents but also the speaker's intention,
3. and, how to choose a set of *surface forms* that the speaker might utter about the predicted concept.

Also, MINDS was concerned with a system to participate in human-machine dialogue. On the other hand, we want to monitor a human-human dialogue. We proposed a dialogue understanding model[7], and a context-sensitive method to predict abstract information about both the intentional and propositional contents of the next utterance[11]. These are our answers to the above problems 1 and 2.

From the point of view of human behavior, a potential approach to selecting the appropriate *surface expression forms* (SEFs) is using *spoken-language* knowledge. In general, when we are talking about a concept X, there are many possible surface expressions and forms to represent X. From a psychological (or psycholinguistic) point of view, Clark[3] pointed out five abstract factors which should be considered in asking *what linguistic devices should speakers use?*. These are: *knowledge of the listener, the cooperative principle, the reality principle, the social context, and the linguistic devices available.* In the computational linguistics area, Appelt[1] has developed a framework to generate a sentence in a context-sensitive way, based on speech act theories. Unfortunately, however, there also remains, as he described as a future study, the problem of choosing a lexically appropriate SEF from among candidates *in a social context*.

This paper describes a context-sensitive framework for selecting an SEF for noun-phrases(NPs). This method is sensitive to both the utterance situation and the history of the dialogue. To do this, first, we analyze the relations between concepts and SEFs, and between applicable situations and contexts, using a corpus of Japanese inquiry dialogues. Then, we make a domain-dependent knowledge source for NP usage, and define rules driven by applicable conditions to determine a set of possible SEFs in the knowledge base. Finally, we give examples of the SEF selection, especially for polite expressions, deictić expressions, and compound NPs, which are common in our target domain, and describe a simple experiment to evaluate using the ATR dialogue database. The result show that the method can choose the contextually correct expression from the speech recognition output candidates, and can be used in the generation module of a spoken-language processing system to generate and determine an appropriate expression under the dialogue situation.

Throughout this paper, all examples are in Japanese and written in *italic*. English translations follow in parentheses. NP denotes a noun phrase, and SEF denotes a *surface expression form*. SEFs are enclosed in double quotation marks and concepts are enclosed in single quotation marks.

2 Dialogue Interpretation and Predicting the Next Utterance

The next utterance can be predicted after understanding the previous utterances, because predicted information must be affected by the dialogue structure. This section briefly describes the model for interpreting a dialogue[7] and the method of predictng the next utterance[11][12].

In the model, an utterance is represented by a predicate form. An typical Japanese sentence, "*Go-juusyo wo onegai-shi-masu.*" (May I have your address?), uttered by the secretariat in a inquiry dialogue, is shown below:

```
(ASK-VALUE s q (address q) (IS (address q) ?val))
```

where constant s denotes the secretariat, q the questioner, address the concept of an address, and the variable, ?val, is the value for the address of q.

The dialogue interpretation model has four types of plan and can interpret input utterances as the dialogue proceeds, using an extended plan inference mechanism[7]. Thus, a dialogue structure can be constructed.

In order to provide contextual information about discourse entities we use *typed variable* notation[2] to describe a discourse entity in a plan schema. Each *type* in this notation corresponds to a particular concept node in the domain-dependent NP knowledge base (described in Section 3.3). The following description is an example of a *Domain Plan* to send something:

```
(Domain-Plan:SEND-SOMETHING
  (HEADER   (SEND ?a:person ?r:person ?s:object))
  (PRECONDITION (KNOW ?a ?d:destination))
  (EFFEECT      (HAS  ?r ?s))
  (CONSTRAINT   (BELONG ?d ?r)))
```

The state of understanding is managed using two pushdown stacks. The **understanding list** stores completed plans as the current understanding state, and the **goal list** maintains incomplete plans as possibilities and expectations for future goals. By referring to the goal list, the next utterance can be predicted on an abstract level as the dialogue proceeds, using the two generalized rules: **expectation** and **preference**[12].

Predicted utterances are represented in the same style as input utterances. As a result, we can predict two types of information, one about the communicative act types and the other about discourse entities in the propositional contents (or in the topic slot) of the next utterance. Information about a discourse entity may appear in the form of an particular expression if it is in a previous utterance that can be related to the current utterance. Otherwise information will be in the form of a *type* representing a particular concept in the related domain plan. We call such information *contextual information* in the task of selecting the constituents of the next utterance.

3 NP Identification Model

3.1 Change to NP Linguistic Expressions

In general, when we are talking about a concept X, there are many possible surface expressions and forms to represent X. In particular, Japanese has several possible SEFs for a given X, one from the Chinese reading and another based on the original Japanese language (e.g. "*okurisaki*" and "*atesaki*" for 'destination' in Fig. 1). In addition, there are particular phenomena of expression variations depending upon the particular dialogue. For example, if a speaker is uttering his/her own address for the concept 'address', he/she will use "*juusyo*"([my] address), e.g. "*Juusyo-wa Oosaka-shi desu.*"(My address is in Osaka-city.). On the other hand, if he/she is uttering the other participant's address, he/she will use "*go-juusyo*"([your] address (polite form)), e.g. "*Go-juusyo-wo onegai-shi-masu.*"(Your address, please?) These facts lead us to implement knowledge sources on such variations(we call them *changes*) in a computational processing system.

Only by filtering using any intra-sentential knowledge sources, several candidates may remain as syntactically and semantically correct sentences. For example, "*gojuu-shichi*"(fifty-seven) sounds like "*go-juusyo*", and the sentence "*Gojuu-shichi-wo onegai-shi-masu.*"(Fifty-seven, please.) is not only well-formed but also correct in a particular context. It is possible to select the correct candidate by referring to both the context and the situation of the ongoing dialogue. Even so, to pick the surface form, we must know why the speaker has used a given expression to represent a concept.

If we can determine how these NPs change, and what effect they have, then we can choose the speech recognition candidates more accurately.

3.2 Analysis of NP Changes

In order to analyze NP changes in a dialogue we inspected 50 dialogues in a corpus. As a result of the analysis, NP changes are categorized into three main classes: 1) **Change by lexical cohesion:** (this class corresponds to *reiteration*[5]), 2) **Change by different viewpoints:** (described in detail in the next paragraph), and 3) **Change by misrecognition**.

There are two aspects of viewpoint, which are *the standpoint of the agent* and *the node of the concept*. In an inquiry dialogue, the standpoints of the agents are always different. Thus, this class has only two subclasses:

2(a) **point keeping:** both agents see the same node of a concept, and this subclass is divided according to the SEF :

  i. **different expression**– e.g. "*watashi*" and "*Yamaoka-san*".

  ii. **addition of prefix**– e.g. "*juusyo*" and "*go-juusyo*".

  iii. **complex**– a mixture of 2(a)i and 2(a)ii,

2(b) **shifting:** the viewpoint of one of the agents shifts from the node of a concept to a node of a related concept, and this is divided into:

  i. **shortening**– e.g. "*Kokusai-kaigi*"(International Conference) and "*kaigi*" (the conference).

  ii. **uniting**– e.g. "*ryousyuu-syo-to saNka-touroku-syo*"(a receipt and an application form) and "*2-syurui-no syorui*"(two types of forms).

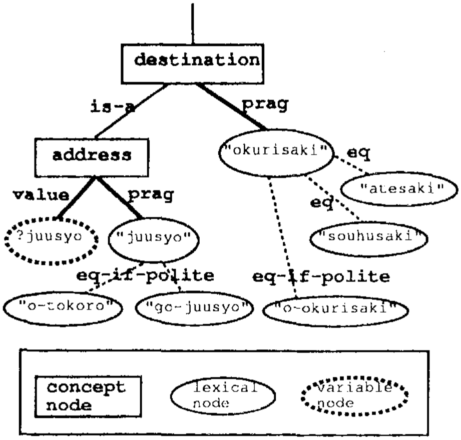

Figure 1: Example of NP knowledge base

iii. **specification**–
e.g. "*niNzuu*"(number of people) and "*saNkaniNzuu*"(number of participants).

3.3 Domain-dependent Knowledge

Configuration: The domain-dependent knowledge base consists of a network of nodes and links. Basic nodes are divided into three types: a **concept node** representing a particular thing or concept retained in human memory, a **lexical node** representing a particular word or phrase used when expressing something, and a **variable node** representing a particular value corresponding to a *valuable* concept, which can have a specified value. A variable node can be instantiated by executing the effect of a completed plan (usually by GET-VALUE-UNIT in *Interaction plan*[11]), so that it can have a particular SEF as the value of the node. For example, "Yamaoka" could be the value of a variable node corresponding to a concept node of 'name' in a sentence like "My name is Yamaoka.".

The following types of links are defined: **is-a link**, representing a superordinate/subordinate relation between two concept nodes, **part-of link**, representing a whole/part relation between two concept nodes, **causal link**, representing a causal relation between two concept nodes, **prag link**, representing a pragmatic relation to connect a particular concept node to a lexical node representing the typical SEFfor the concept, **value link**, representing an instance value relation between a particular valuable concept node and a variable node which has been bound to the SEF of its value, and **eq link**, representing the same meaning between two lexical nodes.

Extension of eq link: In order to make the knowledge base sensitive to the changes considered in Section 3.2, the **eq link** is extended. This lets us to add applicable conditions to **eq links** as sub-types of the link. Applicable conditions are defined based on classes of the categorization in 3.2. For example, if one lexical node is a *polite* SEF of another, the two lexical nodes can be connected with an **eq-if-polite link**, e.g. "*juusyo*" and "*go-juusyo*"(see Fig. 1).

4 Selection Strategy

In the dialogue, a speaker chooses an expression according to the situation, the preceding context, and his/her beliefs. Assuming that the system has recognized such conditions, we can efficiently choose the correct speech recognition candidate by searching the SEFs that are appropriate under the conditions.

4.1 Rules of Applicable Conditions

Here, two terms are defined for explanation:

seed: if the predicted contextual information is bound to a particular SEF, then the seed of the contextual information is the SEF, otherwise the seed is the value of the lexical node linked by the **prag link** to the concept node corresponding to the contextual information,

preferable set: a set of SEFs derived from a seed by an applicable rule, which then takes first priority for selecting the candidate.

The **basic rule** for making a preferable set is: collect the SEFs by following the **eq link** from the seed. Because in this paper we are focusing on dialogue situations and contexts rather than the speaker's beliefs, we only cover rules regarding **changes by different viewpoints**.

For a predicted contextual information I , considering the dialogue situations in Class 2(a):

1. if I is in the territory of information of the other agent, then make a preferable set by following the **eq-if-polite link** from the seed,

additionally, considering the preceding context:

2. if I has an antecedent which denotes the status of the other agent, i.e., there is an instantiated variable node corresponding to I , then replace the seed with the antecedent, i.e., the SEF of the variable node, and make a preferable set by following the **eq-if-polite link** from the seed.

Considering the contexts in Class 2(b):

3. if I is a compound noun, (it's obviously the antecedent) then shift the seed to the concept one-level up [1], and make a preferable set using the basic rule,
4. if I includes two or more concepts or SEFs and there is a concept node which is the upper node of both of these concepts, then shift the seed to the upper concept
and make a preferable set using the basic rule [2],

[1] Precisely, shifting a seed to a concept means an operation to replace the seed with the SEF of the lexical node followed by the **prag link** from the concept node.

[2] In this case an auxiliary word is usually added.

In daily dialogue, speakers apply combinations of the above rules and other rules, but in this study we are concentrating on simpler cases.

4.2 Selection Algorithm

Our ultimate goal is to select the correct speech recognition candidate from the predicted contextual information. An algorithm to do this is roughly defined by following the three steps:

1. provide contextual information,
2. make a preferable set from 1 by the rules,
3. compare speech recognition outputs with 2,
   and if an equivalent is found
   then pick it as the appropriate candidate,
   else goto 2.

Steps 1 and 2 above are backtracking points. For details of Step 1, see [11],[12]. Further large-scale experiments may determine heuristically how many times Step 2 should be iterated.

5 Examples and Evaluation

In this section, we examine some polite expressions and compound NPs that are common in telephone inquiry dialogues.

5.1 Polite expressions

An example of the process for detecting the appropriate SEF given a polite expression, is shown through the following subdialogue, focusing on discourse entities.

(u1) *Q*: *Touroku-youshi-wo okutte-kudasai.*
(Please send me a registration from.)
(u2) *S*: *Go-jusyo-wo onegai-shi-masu.*
(May I have your address?)
(u3) *Q*: *Juusyo-wa Osaka-shi ...... desu.*
(My address is Osaka-city ....)

where agent *Q* is the questioner and *S* is the secretariat.

This example can be recognized in the send-somthing domain plan (in Section 2). First **u1** is recognized and understood as an utterance which introduces the domain plan. Then, for the next utterance by *S* (**u2**), since the system does not hold the statement that *S* knows where to send a form, e.g. the value of *Q*'s address, an utterance requesting the value of the destination is first predicted, and contextual information about the 'destination' concept can be provided (Step 1).

Next, due to the constraint in the plan, Rule 1 is applied to the contextual information. Then, the preferable set of SEFs is derived by the rule(Step 2).

Although the first Step 3 fails because "*go-juusyo*" is not the exact polite form of the first seed "*okurisaki*"(destination), the second time it picks "*go-juusyo*" as the appropriate SEF because one of the lower concepts, 'address', can be the next seed. On the other hand, when processing **u3**, the set of polite forms for the 'address' is not preferred.

Table 1: Result of Database Inspection

Comm. Act Type	Speaker	SEF Type	Number
OFFER-ACTION	SP	polite	25
		normal	4
	HR	normal	22
REQUEST-ACTION	SP	normal	2
	HR	polite	4
		normal	3
CONFIRM-ACTION	HR	polite	3
		normal	5

* The communicative act type in the first column is the type of the utterance "to send". The speaker in the second column is the speaker of the target SEF, with SP indicating the speaker in the first column, and HR, hearer.

Evaluation: We evaluated this method by inspecting SEFs for 'destination' in the ATR dialogue database[4]. The target corpus, whose topic is "Conference registration", has 85 conversations, 1956 utterance units, and 3085 sentences. Moreover, the target expressions are restricted to those uttered in a segment of the send-something domain plan. The evaluation was done in the following way:

1. Retrieve sentences which have the verb "*okuru*" (to send) or synonymous verbs as the main verb of the sentence (161 sentences).
   Then, output the utterance unit together with the next utterance unit (161 pairs).
2. Pick the pairs in which there is a expression about 'destination' (43 pairs).
   Filter by the send-something domain plan, and those pairs that are not recognized are eliminated (32 pairs remain).
3. Classify the target expressions (68 expressions) into the other's territory (32 polite and 12 normal) and the speaker's (24 normal).

The results are shown in the Table 1. This inspection shows that in our target domain, the framework described in the paper is useful for selecting a surface expression that is appropriate in the dialogue situation.

Example (Vocative): Consider the subdialogue that follows the above subdialogue:

(u4) *Q*: *Namae-wa Suzuki-Mayumi-desu.*
(My name is Mayumi Suzuki.)
(u5) *S*: *Suzuki-Mayumi-sama-desu-ne.*
(Ms. Mayumi Suzuki, correct?)

After recognizing **u4** by the same interaction plan as the first example, a variable node corresponding to Q's name is instantiated and bound to "Suzuki Mayumi". Then, for the next utterance by S (**u5**), we can predict the confirmation utterance including the contextual information about Q's name as a discourse entity. Consequently, we can select the SEF "*Suzuki-Mayumi-sama*"(polite form) by the contextual information and the applicable rule 2.

5.2 Compound NPs

Compound NPs can roughly be classified into proper NPs and common NPs. Predicting SEFs from a common NP is usually done by shifting the seed to the upper level (by Rule 3). For example;

(u6) *S*: *Touroku-youshi-wa o-michi-desyou-ka?*
(Do you have an application form?)
(u7) *Q*: *Mada-desu.*
(Not yet.)
(u8) *Youshi-wo okutte-kudasai.*
(Please send me a form.)

In this example, **u6** instantiates the send-something domain plan by the effect chain[7]. Then, since we know from **u7** that the effect (the goal of this subdialogue) is not satisfied, we can predict that the next utterance by Q (**u8**) may concern introducing the action to send a form, and it includes contextual information about 'application form'. In the knowledge base, 'form' is the concept node just above 'application form'. Consequently, by applicable rule 3, we select "*Youshi*"(form) directly.

On the other hand, predicting SEFs for proper NPs requires another rule to create the domain-dependent knowledge base for shortening. Here, we use the dependency relationships within NP[9] to abbreviate a proper compound NP. For example, applying this rule to a proper compound NP "*Kyoto-Kokusai-Kaigijou*"(Kyoto International Conference Center), we get a preferable set of SEFs including "*Kyoto-Kaigijou*"(Kyoto Conference Center) and "*Kokusai-Kaigijou*"(International Conference Center), in addition to the basic upper SEF "*Kaigijou*"(conference center). Consequently, we can select "*Kokusai-kaigijou*" in **u10** in the following subdialogue with a take-transportation domain plan;

(u9) *S*: *Kyoto-kokusai-kaigijou-ewa basu-ga riyou-deki-masu.*
(There is a bus that goes to the Kyoto International Conference Center.)
(u10) *Q*: *Kokusai-kaigijou-made ikura-desu-ka?*
(How much is it to the International Conference Center?)

At the moment, we define a **short link** to connect lexical nodes created by abbreviation rules to the proper compound NP that instantiates a variable node.

6 Conclusion

This paper has proposed a context-sensitive method of predicting NPs in the next utterance of telephone inquiry dialogues. Abstract information about the constituents of the next utterance can be predicted based on the dialogue interpreting model. Then, domain-dependent knowledge for NP usage was developed based on an extended NP identification model. The knowledge base is characterized by its ability to derive the set of possible surface expression forms from the predicted contextual information. We define rules for applicable conditions, particularly in polite Japanese, based on an analysis of NP changes. Finally, using the above two mechanisms a strategy was proposed for selecting the appropriate surface expression form representing the predicted concept in a context-sensitive way.

In the future, we plan to integrate this method with a method of predicting expressions of the speaker's intention, tp form a complete system. It is also vital to make the method more powerful, so it can automatically construct the domain-dependent knowledge base from thesauri and/or corpora of the domain, and can model and recognize various dialogue situations.

Acknowledgements

The authors would like to thank Dr. Kurematsu, President of ATR Interpreting Telephony Labs., and other colleagues for their encouragement and thought-provoking discussion.

References

[1] Douglas E. Appelt. *Planning English Sentences.* Studies in Natural Language Processing. Cambridge University Press, 1985.

[2] Eugene Charniak. Motivation analysis, abductive unification, and nonmonotonic equality. *Artificial Intelligence*, 34:275–295, 1988.

[3] Herbert H. Clark and Eve V. Clark. *Psychology and Language –An Introduction to Psycholinguistics–*, chapter 6, pages 223–258. Harcourt Brace Jovanovich, 1977.

[4] Terumasa Ehara, Kentaro Ogura, and Tsuyoshi Morimoto. ATR dialogue database. In *Proceedings of ICSLP'90*, pages 1093–1096, November 1990.

[5] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*, chapter 6, pages 274–292. LONGMAN, 1976.

[6] Alexander G. Hauptmann, Sheryl R. Young, and Wayne H. Ward. Using dialog-level knowledge sources to improve speech recognition. In *Proceedings of AAAI'88*, pages 729–733, 1988.

[7] Hitoshi Iida, Takayuki Yamaoka, and Hidekazu Arita. Three typed pragmatics for dialogue structure analysis. In *Proceedings of COLING'90*, pages 370–372, August 1990.

[8] Akio Kamio. *Proximal and Distal Inforamtion: A Theory of Territory of Information in English and Japanese.* PhD thesis, University of Tsukuba, March 1986.

[9] Masahiro Miyazaki. Automatic segmentation method for compound words using semantic dependency relationships between words. *Journal of Information Processing Society of Japan*, 25(6):970–979, 1984. in Japanese.

[10] Izuru Nogaito and Hitoshi Iida. Noun phrase identification in dialogue and its application. In *Proceedings of 2nd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, June 1988.

[11] Takayuki Yamaoka and Hitoshi Iida. A method to predict the next utterance using a four-layered plan recognition model. In *Proceedings of ECAI'90*, pages 726–731, August 1990.

[12] Takayuki Yamaoka and Hitoshi Iida. Dialogue interpretation model and its application to next utterance prediction for spoken language processing. In *Proceedings of Eurospeech'91*, September 1991.