Implementing a Japanese Semantic Parser Based on Glue Approach

Hiroshi Umemoto

Fuji Xerox Co., Ltd.
430 Sakai, Nakaimachi, Ashigarakami-gun, Kanagawa 259-0157, Japan,
Hiroshi.Umemoto@fujixerox.co.jp

Abstract. This paper describes the implementation of a Japanese semantic parser based on glue approach. The parser is designed as domain-independent, and produces fully scoped higher-order intensional logical expressions, coping with semantically ambiguous sentences without storage mechanism. It is constructed from an English semantic parser on top of Lexical-Functional Grammar (LFG), and it attains broad coverage through relatively little construction effort, thanks to the parallelism of the LFG grammars. I outline the parser, and I present the analyses of Japanese idiosyncratic expressions including floated numerical quantifiers, showing the distinct readings of distributive and cumulative, as well as a double-subject construction and focus particles. I also explain the analyses of expressions that are syntactically parallel but semantically distinct, such as relative tense in subordinate clauses.

Keywords: formal semantics, syntax-semantics interface, linear logic, semantic composition

1 Introduction

The glue approach to semantic interpretation (glue semantics) provides the syntax-semantics interface where its semantic composition is represented as a linear logic derivation [1, 2]. Glue semantics realizes semantic ambiguity with multiple proofs from the same set of premises corresponding to syntactic items in a sentence, and then it requires no special machineries such as storages. Employing linear logic, it guarantees the semantic completeness and coherence of its results. In other words, all of the requirements of the premises are satisfied and no unused premises remain in linear logic derivations. In addition, glue semantics is not restricted to any specific formalism although it was primarily developed for Lexical-Functional Grammar (LFG), and it can be used with both various syntactic formalisms and meaning representations [3].

I will present the implementation of a Japanese semantic parser based on glue approach. The implementation of glue semantics for English was constructed and incorporated into real-world application systems [4, 5]. However, the theory has not been applied to other languages than English in broad coverage. The Japanese parser is constructed in the same framework as the English parser. Both parsers are designed as domain-dependent, and they produce fully scoped higher-order intensional logical expressions of the kind familiar to traditional formal semanticists as the meanings of sentences. The Japanese parser, as well as the English counterpart, aims to cover wide-ranging real texts, and is built on top of an industrial broad-coverage Japanese grammar based on LFG [6]. The Japanese grammar has been developed through the Parallel Grammar (ParGram) project. The ParGram project uses the Xerox Linguistic Environment (XLE), an efficient parser and grammar development platform based on LFG, and aims to test the formalism for its universality and coverage limitations and see how far parallelism can be maintained across languages [7].

2 Outline of the Parser

The interpretation of glue semantics is comprised of two phases. The first phase is to give rise to meaning constructors, each of which consists of a meaning term and a logical formula, corresponding to the

items in each syntactic analysis of a given sentence. The second phase is to assemble the meaning constructors to derive the meaning expressions of the whole sentence.

The first phase is implemented as description by analysis [2], in which the premises are obtained by analysis of f(unctional)-structure on a rule-by-rule basis. The mapping rules are defined in semantic lexicons that do not access to syntactic lexicons directly, and most of the rules are straightforwardly applicable to Japanese analyses with small modification thanks to the parallelism of the grammars. The derivation machinery in the second phase is commonly used over languages. Semantic composition is handled by linear logic proofs on the meaning constructors, guided by the logical formulas, and it produces all the possible meaning expressions of the syntactic analysis, combining the meaning terms by means of the Curry-Howard isomorphism. Semantic ambiguity, such as scope ambiguity, corresponds to multiple proofs from the same set of meaning constructors [8].

An interpreter for glue semantics is built in the XLE, and it provides a language-independent framework for the theory [1]. Semantic lexicons define mapping rules from f-structure entries into meaning constructors, and are language-dependent. Each of the mapping rules consists of a conditional part and a body part. The former can check morphological labels, predicate forms and f-structure relations regarding to an f-structure entry, and the latter gives meaning constructors and scope constraints if the former is satisfied.

The semantic lexicons for Japanese I implemented consists of 258 mapping rules. Among them, 109 rules are newly created for Japanese semantics, and not exists in the original English lexicons. The newly created rules correspond to verb frames, for examples, with double obliques, predicative adjectives, absolute or relative tense, focus particles, adnominals, distributive readings, modal expressions, and others.

72 rules are used in the Japanese lexicons as well as the English ones, and 77 rules are not changed in their meaning constructor parts but slightly modified in their f-structure conditions for adapting to the Japanese LFG outputs. Incidentally, 104 rules that exist in the English lexicons are removed.

3 Analyses of Japanese Sentences

3.1 Syntactically Parallel and Semantically Ambiguous Sentence

Mapping rules for English are basically straightforwardly applicable to Japanese parallel expressions with small modification. Consider the following sentence:

(1) dono gakusei-mo tukue-o hakonda.
every student even desk-ACC carry-PAST
`Every student carried a desk/desks.'

The sentence (1) has only one syntactic analysis. However, the sentence has at least two possible readings semantically. One of the readings is that the accusative noun representing *desk* takes a narrower scope, and the meaning of the sentence is that there exists a possibly distinct desk for each student. The other reading is the accusative noun takes a wider scope, and the meaning of the sentence is there exists only one desk for all students.

The f-structure corresponding to the sentence (1) is shown in **Fig. 1**. The f-structure contains nested sub f-structures, and each f-structure is tagged by a number at the left-hand side of its open bracket with a colon.

The f-structure shown in **Fig. 1** give rise to the meaning constructors shown in **Fig. 2**.

[1] The glue interpreter and semantic lexicons for English were created by Dick Crouch, Ash Asudeh and John Fry.

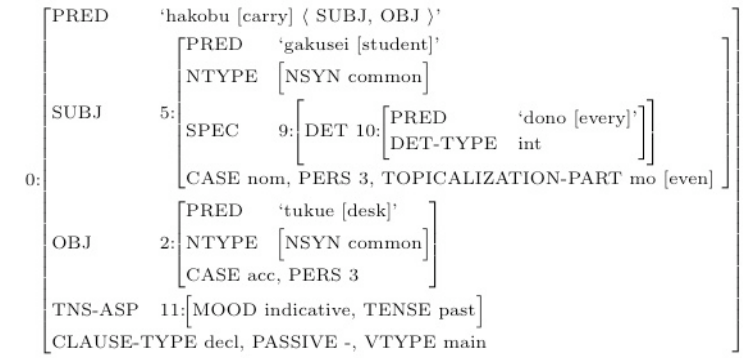

Fig. 1. F-structure corresponding to the sentence (1).

$$
\begin{aligned}
&[\textsc{past}]\ 1\ \lambda A.\lambda B.\lambda C.\texttt{[and,[C,B],[A,B]]}: \\
&\qquad (\textsc{var}_e(\mathbf{11}) \multimap \mathbf{11}_t) \multimap \textsc{evar}_e(\mathbf{0}) \multimap (\textsc{evar}_e(\mathbf{0}) \multimap \mathbf{0}_t) \multimap \mathbf{0}_t \\
&\qquad 2\ \lambda A.\texttt{[past,A]} : \textsc{var}_e(\mathbf{11}) \multimap \mathbf{11}_t \\
&[\textsc{event}]\ 3\ \lambda A.\texttt{[quant,exists,sg,B\[and,[event,B],cxr(C,B,event)],A]}: \\
&\qquad (\textsc{evar}_e(\mathbf{0}) \multimap \mathbf{0}_t) \multimap \mathbf{0}_t \\
&[\textsc{desk}]\ 4\ \lambda A.\lambda B.\texttt{[quant,exists,sg\_pl,C\[and,[A,C],cxr(D,C,barenoun)],B]}: \\
&\qquad (\textsc{var}_e(\mathbf{2}) \multimap \textsc{restr}_t(\mathbf{2})) \multimap (\forall \mathbf{H}_t.(\mathbf{2}_t \multimap \mathbf{H}_t) \multimap \mathbf{H}_t) \\
&\qquad 5\ \lambda A.\texttt{[[noun,tukue(desk)],A]} : \textsc{var}_e(\mathbf{2}) \multimap \textsc{restr}_t(\mathbf{2}) \\
&[\textsc{every}]\ 6\ \lambda A.\lambda B.\texttt{[quant,forall,sg,C\[and,[A,C],cxr(D,C,int(every))],B]}: \\
&\qquad (\textsc{var}_e(\mathbf{5}) \multimap \textsc{restr}_t(\mathbf{5})) \multimap (\forall \mathbf{H}_t.(\mathbf{5}_t \multimap \mathbf{H}_t) \multimap \mathbf{H}_t) \\
&[\textsc{student}]\ 7\ \lambda A.\texttt{[[noun,gakusei(student)],A]} : \textsc{var}_e(\mathbf{5}) \multimap \textsc{restr}_t(\mathbf{5}) \\
&[\textsc{carry}]\ 8\ \lambda A.\lambda B.\lambda C.\texttt{[and,[[verb,hakobu(carry)],A],} \\
&\qquad \texttt{[and,[subcat,A,'V-SUBJ-OBJ'],[and,[xle\_gf,A,B,subj],} \\
&\qquad\ \texttt{[xle\_gf,A,C,obj]]]]} : \textsc{evar}_e(\mathbf{0}) \multimap \mathbf{5}_t \multimap \mathbf{2}_t \multimap \mathbf{0}_t
\end{aligned}
$$

Fig. 2. Meaning constructors corresponding to the sentence (1).

Columns in the above table show labels, sequence numbers, and meaning constructors respectively. The left-hand side of a colon symbol represents a meaning expression, and the right-hand side provides a logical formula. The body of a meaning expression is shown in typewriter fonts, in which a capital letter represents a variable and the backslash symbol stands for lambda abstraction. A bold number in a logical formula represents the meaning of the f-structure tagged by the number. EVAR, VAR and RESTR are features of semantic structure standing for event variable, variable and restriction respectively. Small characters *e* and *t* on the lower right of a term stand for the type of entity and that of truth-value respectively. A symbol consisting of a bar and a circle stands for implication in linear logic.

In meaning expressions, a generalized quantifier is expressed in the Prolog list notation, where its elements are the identifier **quant**, quantifier type **exists** or **forall**, cardinality such as **sg** that stands for singular, a restriction part, and a body part. The function **cxr** is placed as a hook to mark contextual dependencies. Subcategorization information is explicitly represented in **subcat** and **xle_gf** entries, and the latter stands for XLE grammatical function.

Common nouns in Japanese are not marked with regard to number. The cardinality of the noun representing *desk* in the sentence is set to **sg_pl**, which does not specify the cardinality. Common nouns are often expressed as bare nouns, and the definiteness of common noun is usually underspecified.

Following two scope constraints are also imposed:

$$\texttt{notscopes}(\textsc{evar}_e(\mathbf{0}), \textsc{restr}_t(\mathbf{2}))$$
$$\texttt{notscopes}(\textsc{evar}_e(\mathbf{0}), \textsc{restr}_t(\mathbf{5}))$$

These constraints mean the event variable of node 0 does not outscope the restrictions of node 2 and

node 5.

All of the semantic expressions corresponding to the f-structure are to be obtained via logical derivation. There are several possible proofs. One proof is to consume meaning constructors in the order of meaning constructors: (1, 2), 3, 8, (4, 5), (6, 7). Meaning constructors enclosed within parentheses are consumed before the resultant formulas are consumed with the right-hand neighbor meaning constructors. This proof derives the following meaning expression corresponding to the narrow reading:

```
[quant,forall,sg,J\[and,[[noun,gakusei(student)],J],cxr(K,J,int(every))],
 L\[quant,exists,sg_pl,M\[and,[[noun,tukue(desk)],M],cxr(N,M,barenoun)],
    O\[quant,exists,sg,P\[and,[event,P],cxr(Q,P,event)],
       R\[and,
          [and,[[verb,hakobu(carry)],R],
           [and,[subcat,R,'V-SUBJ-OBJ'],
            [and,[xle_gf,R,L,subj],[xle_gf,R,O,obj]]]],
          [past,R]]
       ]
    ]
 ]
```

Another proof is to consume meaning constructors in the order of (1, 2), 3, 8, (6, 7), (4, 5). This proof derives the meaning expression corresponding to the wide reading. Other distinct proofs might be considered, for example (4, 5), 8, (6, 7), (1, 2), 3 or (6, 7), 8, (4, 5), (1, 2), 3, but the scope constraints prevent them.

3.2 Japanese Idiosyncratic Expressions

Floated Numeral Quantifiers

Subject-oriented floated quantifiers syntactically combine with VP, whereas object-oriented floated quantifiers combine with the verb only. The former constructions do not allow a collective reading, but a distributive one. There is a mapping from events to individuals, and measure functions indirectly measure events by measuring individuals related to the events [9]. The following is a sample sentence containing floated numeral quantifiers [2] :

(2) gakusei-ga 3-nin biiru-o 6-hai nonda.
student-NOM 3-CLS beer-ACC 6-CLS drink-PAST
`Three students drank six glasses of beer.'

I divided classifiers into two types: one is specifying the cardinality of a countable referent, and the other is measuring an uncountable referent. If a subject-oriented floated quantifier has a countable referent, there are distributive multiple events, each of which is corresponding to one of the elements represented by the host noun. I introduced two features of semantic structure: a distributive element of the host noun and a distributive event variable, standing for D and DVAR respectively.

The sample sentence (2) has at least two possible readings: a cumulative reading and a distributive reading in which the scope of *3-nin* contains the denotation of *6-hai*. The total number of glasses of beer that are drunk is six in the former reading and eighteen in the latter [10]. From the meaning constructors and scope constraints corresponding to the sentence (2), seven analyses are obtained, which contains both readings. A final meaning expression corresponding to the cumulative reading is shown below:

[2] This example is taken from [10] with modification.

```
[quant,exists,3,
 A\[and,[and,[[noun,gakusei(student)],A],[classifier,nin,A]],cxr(B,A,numeral)],
 C\[quant,exists,pl,D\[and,[event,D],cxr(E,D,event)],
    F\[quant,exists,6,
       G\[and,[and,[[noun,biiru(beer)],G],[classifier,hai,G]],cxr(H,G,numeral)],
       I\[quant,forall,sg,J\[member,J,C],
          K\[quant,exists,sg,L\[evmember,L,F],
             M\[and,
                [and,
                 [and,
                  [and,[[verb,nomu(drink)],M],
                   [and,[subcat,M,'V-SUBJ-OBJ'],
                    [and,[xle_gf,M,K,subj],[xle_gf,M,I,obj]]]],
                  [past,M]],
                 [adverbial,measure,M]],
                [adverbial,measure,M]]
             ]
          ]
       ]
    ]
 ]
```

Double-Subject Construction

Sentences where adjectives dominate two surface subjective cases are common in Japanese, which are called double-subject construction sentences [12]. The basic sentence pattern of the double-subject construction is `X *wa* Y *ga* -- predicate', where X and Y are NPs [11]. Here is a sample sentence:

(3) zoo-wa subete hana-ga nagai.
elephant-TOP every nose-NOM long-PRES
`Every elephant has a long trunk.'

The double-subject construction is classified into several types, and in the case of the above sentence, the topic marker *wa* is a proxy for the particle *no* representing a noun modifier [12]. In other words, the subject word *hana* and the topicalized word *zoo* have a part-whole relation. The meaning expression of the sentence is as follows.

```
[quant,exists,sg_pl,A\[and,[[noun,zoo(elephant)],A],cxr(B,A,barenoun)],
 C\[quant,exists,sg_pl,
    D\[and,[and,[[noun,hana(nose)],D],[of_topic,C,D]],cxr(E,D,barenoun)],
    F\[quant,exists,sg,G\[and,[event,G],cxr(H,G,event)],
       I\[and,
          [and,
           [and,[[adjective,nagai(long)],I],
            [and,[subcat,I,'A-SUBJ'],[xle_gf,I,F,subj]]],
           [pres,I]],
          [postposition,topic,wa,I]]
       ]
    ]
 ]
```

Focus Particles

Japanese focus particles have ambiguities, but [13] reports the two level rules for morphemes and grammatical functions disambiguate the particles *made*, *nado* and *dake* in high precision. Consider the following examples:

(4) Taro-dake-ga kasikoi.
Taro only-NOM clever-PRES.
`Only Taro is clever.'

(5) Taro-dake-ni kasikoi.
Taro because-ADJUNCT clever-PRES
`Taro is clever because he is Taro (known to be clever).'

The particle *dake* tends to show the function of restricting its PRED(icate) noun if it precedes a case marker. On the other hand, it tends to present the function of reasoning if it is included in an adjunct [13]. The meaning expressions of the above sentences are shown below:

```
[[dake,only],name(sigma(3),A\[and,[name,taro,A],[name_type,A,name]]),
 B\[quant,exists,sg,C\[and,[event,C],cxr(D,C,event)],
    E\[and,
       [and,[[adjective,kasikoi(clever)],E],
        [and,[subcat,E,'A-SUBJ'],[xle_gf,E,B,subj]]],
       [pres,E]]
    ]
 ]

[quant,exists,sg,A\[and,[nullPro,A],cxr(B,A,nullPro)],
 C\[quant,exists,sg,D\[and,[event,D],cxr(E,D,event)],
    F\[and,
       [and,
        [and,[[adjective,kasikoi(clever)],F],
         [and,[subcat,F,'A-SUBJ'],[xle_gf,F,C,subj]]],
        [pres,F]],
       [postposition,
        G\[[dake,reason],
           name(sigma(4),H\[and,[name,taro,H],[name_type,H,name]]),G],
        ni,F]]
    ]
 ]
```

Syntactically Parallel but Semantically Distinct Expressions

There are exceptions for applying mapping rules for English to Japanese syntactically parallel expressions. One of them is relative tense in subordinate clauses. It is claimed that Japanese has a relative tense system, and a tense morpheme is interpreted in relation to the tense that locally c-commands it [14]. Consider the following sentence:

(6) watasi-ga nihon-e iku toki-ni purezento-o katta.
I-NOM Japan-to go-PRES time-DAT present-ACC buy-PAST
`When/Before I went to Japan, I bought a present/presents.'

Regardless of whether the main clause is in the present or the past, the subordinate clause adjoined to *toki* takes the present tense if the action in the clause has not been completed before the action expressed in the main clause takes place. If, however, the action in the *toki* clause has been completed before the action in the main clause takes place, the *toki* clause takes the past tense. [11]

(7) watasi-ga nihon-e itta toki-ni purezento-o katta.
I-NOM Japan-to go-PAST time-DAT present-ACC buy-PAST
`When/After I went to Japan, I bought a present (in Japan).'

If the *toki* clause takes the stative or ongoing verb, the eventualities of both clauses take place simul-

taneously. Therefore, in the case of the former sentence I specify that the eventuality of the main clause does not always follow that of the *toki* clause but *not-precedes* it.

However, tense morphemes in relative clauses are not always relative. One of the examples where tense morphemes in relative clauses are absolute is the following:

(8) jisatu-sita hito-ga takusii-ni notta.
suicide-PAST person-NOM taxi-DAT ride-PAST
`A person who killed oneself took a taxi.'

It is impossible to interpret the above sentence that the suiciding event preceded the riding event, because a dead person cannot take a taxi. According to the context, a tense morpheme in a relative clause should be treated as an absolute one. The event of the relative clause is a fact and happened in the past if both clauses take the past tense, although it does not always precede the event of the main clause. Symmetrically, it happens or will happen in the future if both clauses take the present tense, although it does not always follow the event of the main clause.

(9) jisatu-suru hito-ga takusii-ni notta.
suicide-PRES person-NOM taxi-DAT ride-PAST
`A person who would kill oneself took a taxi.'

3.4 Coverage

A parsing experiment was conducted against a portion of EDR Japanese corpus [15] that contains more than 200,000 sentences collected from newspapers, magazines and dictionaries. 1,000 sentences were randomly selected from the corpus, and the average number of words per sentence is 21. The Japanese LFG parsed 942 sentences (857 full parses and 85 fragment parses that were parsed as well-formed chunks specified by a fragment grammar [4, 6]), and failed to analyze 58 sentences within a time limit. 675 sentences received a full semantic analysis for at least one possible parse, and the ratio between full semantic analyses and syntactic parses consisting of both full and fragment parses is 71.7%. The ratio of full semantic analyses to full syntactic parses is 77.9% (668 full semantic analyses for 857 full syntactic parses).

Subject	Sentences	Semantic analyses	Coverage (%)
Total	1,000	675	67.5
LFG parses	942		71.7
LFG full parses only	857	668	77.9
LFG fragment parses only	85	7	8.2
LFG failure	58	0	0

Table 1. Result of the coverage experiment using EDR Japanese corpus 1,000 sentences.

4 Conclusion

I presented the implementation of a Japanese semantic parser based on glue approach. First, I showed that a Japanese sentence that is syntactically parallel to the English translation is analyzed as well as the English counterpart by the parser. Second, I explained the analyses of Japanese idiosyncratic expressions such as floated numerical quantifiers, a double-subject construction and focus particles. Third, I mentioned the analyses of relative tense in subordinate clauses that are syntactically parallel but semantically distinct. The parser attains broad coverage through relatively little construction effort including porting English semantic lexicons, thanks to the parallelism of the LFG grammars.

Acknowledgements. I gratefully acknowledge the support of Dick Crouch at Palo Alto Research Center Inc. for helping me to understand the theory and the system of glue semantics.

References

1. Dalrymple, M., ed.: Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach. The MIT Press, Cambridge, MA (1999)
2. Dalrymple, M., ed.: Lexical Functional Grammar. Volume 34 of Syntax and Semantics. Academic Press (2001)
3. Asudeh, A., Crouch, R.: Glue semantics for HPSG. In: Proceedings of the 8th Intl. HPSG Conference. (2001)
4. Crouch, D., Condoravdi, C., Stolle, R., King, T., de Paiva, V., Everett, J.O., Bobrow, D.G.: Scalability of redundancy detection in focused document collections. In: Proceedings First International Workshop on Scalable Natural Language Understanding (SCANALU-2002). (2002)
5. Crouch, D.: Packed rewriting for mapping semantics to KR. In: Proceedings of the 6th International Workshop on Computational Semantics. (2005)
6. Masuichi, H., Ohkuma, T.: Constructing a practical Japanese parser based on Lexical-Functional Grammar. Journal of Natural Language Processing 10(2) (2003) 79-109 (in Japanese).
7. Butt, M., Dyvik, H., King, T.H., Masuichi, H., Rohrer, C.: The parallel grammar project. In: Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation. (2002) 1-7
8. Asudeh, A.: Resumption as Resource Management. PhD thesis, Stanford University (2004)
9. Nakanishi, K.: The semantics of measure phrases. In: the Proceedings of the 33rd Conference of the North East Linguistic Society (NELS 33). (2003) 225-244
10. Gunji, T., Hasida, K.: Measurement and quantification. In Gunji, T., Hasida, K., eds.: Topics in Constraint-Based Grammar of Japanese. Kluwer Academic Publishers, Dordrecht (1998)
11. Storm, H.: A Handbook of Japanese Grammar. Volume 18 of LINCOM Handbooks in Linguistics. LINCOM GmbH (2003)
12. Oku, M.: Analysing Japanese double-subject construction having an adjective predicate. In: COLING-96. (1996) 865-870
13. Ohkuma, T., Masuichi, H., Yoshioka, T.: Disambiguation of Japanese focus particles by using lexical functional grammar. Journal of Natural Language Processing 13(1) (2006) 27-52 (in Japanese).
14. Ogihara, T.: Tense and aspect. In Tsujimura, N., ed.: The Handbook of Japanese Linguistics. Blackwell Publishers (1999) 326-348
15. Japan Electronic Dictionary Research Institute, Ltd.: EDR Electronic Dictionary Technical Guide. (1996)