Multiple argument ellipses resolution in Japanese

Shigeko Nariyama

MIALS, The University of Melbourne
Parkville Victoria Australia 3010
shigeko@unimelb.edu.au

Abstract

Some Japanese clauses contain more than one argument ellipsis, and yet this fact has not adequately been accounted for in the study of ellipsis resolution in the current literature, which predominantly focus resolving one ellipsis per sentence. This paper proposes a method using a "salient referent list", which identifies the referents of such multiple argument ellipses as well as offers ellipsis resolution as a whole by considering contextual information.

Key words

Ellipsis resolution, multiple argument ellipses, salient referent list

1 Introduction

It has been widely recognised and has been a challenge in Japanese-to-English machine translation that Japanese frequently unexpresses nominal arguments, such as the subject and the object, which must be identified and made explicit in order to be translated into grammatical English. So far most attempts have been made on resolving only one ellipsis in a clause (not a sentence), predominantly of the subject ellipsis. However, some clauses contain multiple ellipses within the clause, e.g. the subject as well as the object.

This paper proposes a method using a "salient referent list" for resolving such multiple argument ellipses by considering contextual information. This single method is not ad hoc, but offers a unified account that resolves not only multiple argument ellipses, but also ellipsis resolution as a whole.

2 Earlier studies

Studies on ellipsis resolution in Japanese have focused to resolve one ellipsis per simplex sentence. For example, Centering Theory (Kameyama 1985, Walker et al. 1994) is confined to assign only one Cb (backward looking center) per sentence, consequently resolving at best one ellipsis. Kameyama (1986) proposed the "property sharing constraint" which is meant to account for this problem. However, as is shown later in this section, some examples show the inadequacy of her proposal.

In addition, Centering Theory is particularly problematic, in that it mostly deals with simplex sentences, despite the fact that complex sentences are reported to comprise 87.5% of Japanese narrative texts (Nariyama 2000). In more recent work, Kameyama (1998) proposed an account for intrasentential Centering by breaking a complex sentence into a hierarchy of center-updating units, i.e. clauses in more general terms. My assumption on her account is to utilise conjunctive particles, by which a hierarchy of center-updating units for each complex sentence is determined. However, this account requires additional convoluted hierarchies and yet its results were shown to be still inadequate by Strube (1998). Even with the potential increase in the accuracy, this method still retains an unsolved problem – resolving non-subject ellipsis. Moreover, it is designed for English complex sentences, and implications for Japanese sentences are not addressed in her work.

A similar account is taken by Nakaiwa who has produced a series of works on Japanese ellipsis resolution (1995, 1998, inter alia), which are said to account for intrasentential ellipsis by means of conjunctive particles. However, conjunctive particles per se can retrieve at best the identity of only a subject ellipsis, in terms of same subject (SS), different subject (DS), or no prediction between any two adjacent clauses.

Hence, given the current understanding of ellipsis resolution, complex sentences with multiple argument ellipses within a clause, such as (1)[1], are not licensed to systematically resolve their identities using the methods devised by Kameyama and Nakaiwa et al.

(1) [*Watasi*$_a$*-wa Goo*$_c$*-no fan de*,]$_1$
I-TopSB Goo-Gen fan be[SS],
[*Goo*$_c$*-ga marason-ni choosensuru node,*]$_2$
Goo-SB marathon-Obl challenge because[DS]
[$\phi_{\mathbf{a}}$ $\phi_{\mathbf{c}}$ *ooensi takatta*]$_3$ *kara da*.$_4$
SB OB cheer wanted because be
"The reason is that I_a am a fan of *Goo*$_c$, and because *Goo*$_c$ was challenging the marathon, (I_a) wanted to cheer for (him$_c$)."
[Taken from *Seikachoo* Newspaper (2.1999)]

[1] Each subordinate clause is indicated by square brackets [ ] with the clause number on the right side. The matrix clause is numbered but not bracketed.

Theoretically speaking, the *wa*-marked (topicalised) subject in Clause 1 should belong to Clause 3, since syntactically it is considered to be preposed to the front of the sentence. However, in practice and more realistically, due to the constraints from short term memory, segments of a sentence are processed as they are produced, so that in this paper the *wa*-marked subject is processed as the subject of Clause 1, which provides the same reading as the method treating it as the subject of Clause 3.

The following abbreviations are used in the examples: ø=ellipsis, DS=different subject, Gen=genitive (possessive), Nomz=nominalizer, IO=indirect object, OB=object, Obl=oblique (arguments other than the subject and the object), SB=subject, SS=same subject, TopSB=topicalised subject.

The clause 3 in (1) contains multiple argument ellipses, i.e. the subject and the object. The ellipted subject is coreferential with the topicalised subject in Clause 1 shown by the subscript 'a', and the ellipted object with the subject 'c' in Clause 2. It is not unequivocal how this reading is licensed using the methods by Kameyama and Nakaiwa et al.

Nakaiwa et al. utilise verbal semantic attributes (using *Goi taikei* Valency dictionary (Ikehara et al. 1997)), whereby the valency information from the verbal semantics *ooensi* 'to cheer' selects the subject and the object, both of which can select human arguments. Given that there are two human arguments '*watasi*' and '*Goo*', it is not obvious how the above reading is reached. Nakaiwa et al. also utilise SS/DS information from the conjunctive particles. Using this information, the DS conjunctive particle in Clause 2 tells us that the ellipted subject is different from the subject in Clause 3. This resolves the problem in this example. Note, however, that the accuracy of the SS/DS reading from the conjunctive particles is reported to be around 60 to 90 percent of the time in corpora (Iwasaki 1993:64, Watanabe 1994:150-2, inter alios; cf. Minami 1974:130). As a matter of fact, this is substantiated in the very same sentence; Clause 1 contains an SS conjunctive particle, in spite of the DS reading specified by the overt subject in Clause 2. Nariyama (2000) has shown that the reading from the conjunctive particles can be overridden by an overt expression of different subject, as in (1), and by the interaction of the subject marking particles: *wa* and *ga*.

Furthermore, Kameyama (1986) proposed the "property sharing constraint", which states that zero pronominal binding is acceptable, if one or more of two properties are shared between the antecedent and the zero pronominal: non/Subject and non/Identification. In other words, ellipsis should apply to the subject if the referent is the subject, and ellipsis should have the "speaker's identification" if the referent does. This does not adequately explain why the object ellipsis above has the subject referent and be perfectly acceptable. Hence, under the Kameyama method, it would have to conclude that resolution of such multiple ellipses requires world knowledge that if *Goo* is running and I am his fan, then I must be the one to cheer for him.

3 Salient referent list

This paper takes an eclectic approach from a number of previous methods and proposes a method using a "salient referent list" for resolving such multiple ellipses. It works to reflect how humans store referential information.

Each sentence contains one or more referents. The subsequent sentences may retain one or more of these referents, some or all of which may be expressed by ellipses, and may also introduce one or more new referents. It is plausible to assume that when processing sentences, addressees store new referents by incorporating them into a pool of old referents from the previous sentences which have been stored in their cognition, and repeat this process as they process each new sentence. The salient referent list does just that. It functions like a memory bank in a cognitive sense, listing overt arguments appearing in the sentence by incorporating arguments that have appeared in the previous sentences. It is this input information which provides cues to resolve ellipses in the sentence, not only for subject ellipsis but also for non-subject ellipses hence resolving multiple argument ellipses.

The salient referent list basically lists all overt arguments which have appeared up to the sentence in question. These overt arguments are listed in the following hierarchical order, called the "salient referent order list",[2] which accords the topicalised subject the highest saliency. In Japanese, the topicalised subject is morphologically differentiated from the non-topicalised subject by the use of different markers: *wa* and *ga* respectively.

Topicalised SB (Global > Local)
> Topicalised SB in quotation
> SB > IO > OB > Obl

Figure 1: Salient referent order list

'Topicalised SB (Global > Local)' is to cover the fact that although when a new topic is introduced, normally the new topic replaces the old one, when there is a global topic (usually the writer or the main topic/protagonist of the text), sometime it is still carried over after a long absence of the mention, while the current (local) topic is still in effect. 'Topicalised SB in quotation' refers to the topic in quotation whereby the scope does not extend beyond the quotation, unlike the normal topic. (See the results in Section 5.)

A salient referent list is created for each new sentence by modifying the one from the preceding sentence. If an argument appears with an identical grammatical relation to another argument already existing in the list, for example, where a subject exists and a new subject appears, the new subject takes its place for reasons of recency, except for topicalised subjects.[3]

4 Creation of salient referent lists and ellipsis resolution

This section explains how salient referent lists are created and used to resolve ellipses including multiple argument ellipses, using a fragment of a text from *Seikachoo* Newspaper (2.1999). Each sentence is numbered, noted as [s1] being the first sentence in the text. Each argument listed in the salient referent list is provided with the grammatical relation, topicality and person/animacy.[4]

---

[2] The salient referent list order was eclectically adapted from the Japanese version of Expected Center Order in Centering Theory (Kameyama 1985), Keenan and Comrie's (1977) noun accessibility hierarchy, Givon's (1979) topicality hierarchy, and Kuno's (1987) thematic hierarchy.

[3] This method of listing only one argument under any one slot of grammatical relation works satisfactory in the texts analysed. However, this needs be further investigated in more texts and larger texts.

[4] For simplicity, this paper notes only the grammatical relation, topicality and person/animacy. However, in practice, other information should be also noted; e.g. detailed semantic attributes of arguments except for first and second person, number, and the in-group/out-group distinction. Due to the limited space, this paper was unable to provide full details of the process (algorithm) and information (grammatical rules)

[s1]
[Watasi$_a$*-wa senshuu no doyoobi hotondo ne nai de]*$_1$
I-TopSB last week of Saturday hardly sleep not and[SS]
$\phi_{\mathbf{a}}$ *terebi*$_b$*-o mi tuzuketa.*$_2$
SB TV-OB watch continued

"Last Saturday, I_a hardly slept, instead (I_a) kept on watching TV_b."

[s1] has only one human argument - the topicalised subject *watasi*, and one inanimate object *terebi*. Hence, the salient referent list (SRL) for [s1] is formulated as follows:

SRL: [s1] {T1$_a$: *watasi* (TopSB; first person) >
T2$_b$: *terebi* (OB; inanimate)}

Each listed argument is given a number, for example, 'T1'. The argument under T1 has the highest saliency and is therefore the best candidate as the referent for the ellipsis; T2 is the next highest, and so forth. They are listed in the salient referent list accordingly. Ellipsis is resolved based on the information in the salient referent list for the sentence where the ellipsis appears. [s1] contains one ellipsis, so that T1 argument is applied as the referent, which is indeed the case. This coreference is indexed by subscript after T1 as 'T1$_a$', which is also coindexed in the text for easy recognition.

The next sentence is denoted as [s2], which is the same as (1).[5]

[s2]
Nazenara, [[watasi$_a$*-wa Goo*$_c$*-no fan de,]*$_1$
Because I-TopSB *Goo*-Gen fan be-and,

[Goo$_c$*-ga marason-ni choosensuru node,]*$_2$
Goo-SB marathon-Obl challenge because[DS]

$\phi_{\mathbf{a}}$ $\phi_{\mathbf{c}}$ *ooen sitakatta]*$_3$ *kara da.*$_4$
SB OB cheer wanted because be

"The reason is that I_a am a fan of Goo_c, and because Goo_c was challenging the marathon, (I_a) wanted to cheer for (him_c)."

The salient referent list needs to be updated with each new sentence, so that each salient referent list also needs to be numbered. In [s2], there are two overt arguments '*watasi*' and '*Goo*'. The referent '*watasi*' appears again with the same function of topicalised subject, so it remains as T1 in the list. The other argument 'c' is a non-topicalised subject, so that it is listed as T2. There is no other argument in [s2], so that the inanimate object argument '*terebi*' from the previous salient referent list is carried over to the salient referent list for [s2]. But this time as T3, because the object is listed lower than the subject in the salient referent order list. Hence, the salient referent list for [s2] is formulated as follows:

SRL: [s2] {T1$_a$: *watasi* (TopSB; first person) >
T2$_c$: *Goo* (SB; third person) >
T3$_b$: *terebi* (OB; inanimate)}

[s2] has multiple ellipses in Clause 3: the subject and the object. Multiple ellipses are also ranked by the same salient referent order list, so that the subject ellipsis is ranked higher than the object ellipsis. The method of multiple argument ellipses resolution works as follows - the T1 argument in the salient referent list is chosen to be the referent for the highest ranked ellipsis in the salient referent order list. Similarly, T2 is selected as the referent for the next highest ellipsis, T3 is for the next highest ellipsis, and so forth. Accordingly, in [s2], the subject ellipsis is ranked higher than the object ellipsis, so that T1 'a' is chosen to be the referent of the subject ellipsis, and T2 'c' as the referent of the object ellipsis. This interpretation, following the proposed method, correctly selects the referents for the multiple ellipses.

Thus, it demonstrates that although the grammatical relations are generally held constant between the referent and the ellipsis, they need not be shared at all times. Hence, this denies the property sharing constraint in favour of the salient referent list.

Due to space limitation, it is necessary to omit [s3] ~ [s5] in order to include [s6] which contains another instance of multiple argument ellipses.

[s6]
Aruhi, [sono hito$_g$*-wa zassi-o mitei tara,]*$_1$

one day that person-TopSB magazine-OB looking when[DS]

[[Hugh$_i$*-san-ga rockclimbing*$_h$*-o si-te]*$_2$
Hugh-Mr-SB rock-climbing-OB do-and[SS]

[$\phi_{\mathbf{i}}$ $\phi_{\mathbf{h}}$ *seikoo siteiru]*$_3$ *koto]*$_4$*] -o* $\phi_{\mathbf{g}}$ *sitta.*$_5$
SB Obl success have been Nomz-OB SB knew

"One day, when the person$_g$ was reading a magazine, (he_g) noticed that $Hugh_i$ attempted $rock\text{-}climbing_h$ and (he_i) succeeded in (it_h)."

SRL: [s6] {T1$_g$: *hito* (TopSB; third person) >
T2$_i$: *Hugh* (SB; third person) >
T3$_h$: *rock-climbing*(OB; inanimate)}

The subject ellipsis in Clause 5 is coreferential with T1. Analogous to [s2], the multiple ellipses are also ranked by the same salient referent order list; i.e. the subject is higher than the oblique. The subject ellipsis in Clause 3, however, is joined by the SS conjunctive particle in the preceding Clause 2, which signals that the subject in Clause 3 is the same as that in Clause 2. Hence, the subject in Clause 2 (i.e. T2 argument) is chosen to be the referent for the subject ellipsis. Consequently, the next referent on the list T3 is chosen for the oblique ellipsis.

These interpretations, following the proposed method, makes a correct selection of the referents for the multiple ellipses as well as the subject ellipses. It demonstrates again that the grammatical relations of referent and ellipsis need not be held constant, and that salient referent list offers the key to resolving ellipses in

---

needed to process ellipsis resolution for clarity and substantiation (see Nariyama 2000).

[5] Although [s1] does not contain multiple ellipses, it had to be explained in order to demonstrate how salience reference lists are created for each consecutive sentence and to lead to the next sentence which does contain multiple ellipses.

Japanese.

5 Results and evaluation

The salient referent list is hand-tested on 7 short essays written by non-professional writers, which eliminate any potential bias caused by individual writing styles and topics. One of these essays is taken from *Seikachoo* Newspaper (2.1999) and the rest from PHP magazines (2.1999). The results are shown in Table 1. There are 210 ellipses.

Texts	T1	T2	T3	T4	T5	T6	T7	Σ
No. of sentence	24	25	9	18	67	9	19	171
No. of ellipsis	39	33	17	25	53	16	27	210
✗ No of incorrect ø	0	2	1	1	15	5	6	30
% of ✗	0	6.0	5.9	4.0	*28.3*	*31.3*	*22.2*	14.3
No. of multiple ø	2	0	1	1	2	0	0	6

Table 1: Effectiveness of Salient Reference List

The focus of attention here is needless to say the effectiveness of the salient reference list. The texts are divided into sharply contrasted two groups in terms of accuracy; the salient reference list is extremely effective for Texts 1~4, but not so for Texts 5~7. There were mainly five factors responsible for the incorrect selections.

The first factor is caused by the lack of precise differentiation of 'global' topic and 'local' topic as to when 'global' topic overrides 'local'. In Texts 5~7, the writers used ellipted 'I' as the global topic at random points.

The second factor is the anomalous use of *ga* (the non-topicalised subject marker) which had scope over to the next sentence, which is normally the function of *wa*. What happened was that *ga* which also has another function of exhaustive listing (focus) overtook the topic marker *wa*. Namely, *wa* would have been used, if it were not focused.

Note that the first and the second factors comprise of 18/30 errors, most of which occurred within the same sentence or in succession, so that the actual occurrence was less than half the times. They are the main triggers for the poor performance for Texts 5~7.

The third occurred when the particle *wa* is used not as the topic marker but as the contrastive marker. The differentiation of the two functions of *wa* is murky and an unresolved issue in linguistics. When *wa* is used as the topic, the ellipsis is coreferential with the *wa*-marked referent. However, when it is used as the contrast, the ellipsis is coreferential with the previous *wa*-marked referent.

The fourth is the problem caused by a part-whole relationship. For example, T1 may list 'John's life', but the ellipsis refers to 'John'.

The fifth is the notorious problem of world knowledge, comprising of 5/30 errors.

With regard to multiple argument ellipsis, there are only 6 cases (5.8%) (c.f. 9.2% found in Nariyama 2000)), and all of them are correctly resolved by the proposed method.[6]

6 Conclusion

This paper demonstrated that the salient referent list stores contextual information and is therefore a promising method for ellipsis resolution. It is particularly robust for multiple ellipses resolution. However, this is a preliminary report based on hand-simulated analysis using short narrative texts. The proposed method requires a large corpus analysis and corpora from difference genres (e.g. newspapers, conversation scripts) to be fully evaluated with consideration to those problems described in Section 5. This is the next step for future research.

References

Givón, T. (1979). *On understanding grammar*. New York: Academic Press

Ikehara, et al. (1997). *Goi-taikei - A Japanese lexicon.* Tokyo: Iwanami Shoten.

Iwasaki, S. (1993). *Subjectivity in grammar and discourse: theoretical considerations and a case study of Japanese spoken discourse*. Amsterdam: John Benjamins

Kameyama, M. (1985). *Zero anaphora: the case of Japanese*. Ph.D Dissertation: Stanford University

Kameyama, M. (1986). A property-sharing constraint in centering. In proceedings of the 24th Annual Meeting of the *Association of Computational Linguistics*, 200-206

Kameyama, M. (1998). Intrasentential Centering: a case study. In M. Walker, K. Joshi and E. Prince (eds.). *Certering theory in discourse*. Oxford: Clarendon Press. 89-112

Keenan, E. & B. Comrie. (1977). 'Noun phrase accessiblility and universal grammar'. *Linguistic Inquiry* 8: 63-99

Kuno, S. (1987). *Functional syntax: anaphora, discourse and empathy*. Chicago: The University of Chicago Press

Minami, F. (1974). *Gendai Nihon-go no kouzou* (Structures of modern Japanese). Tokyo:Taishukan

Nakaiwa, H. (1998). *A study on resolving Japanese zero pronouns in machine translation*. ms

Nakaiwa, H. et al. (1995). Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints. *AAAI '95 Spring Symposium*. 99-105

Nariyama, S. (2000). *Referent identification for ellipted arguments in Japanese.* Ph.D Dissertation: University of Melbourne

Strube, M. (1998). Never look back: An alternative to Centering. In proceedings of the 36th Annual Meeting of the *ACL*: 1251-1257

Watanabe, Y. (1994). Clause-chaining, switch-reference and action/event continuity in Japanese discourse: The case of *TE, TO* and zero-conjunction. *Studies in Language* 18(1): 127-203

[6] In addition, the proposed method also resolved other problematic issues: cataphora and the reflexive pronoun *zibun* which does not differentiate person and gender.

Walker, et al. (1994). Japanese discourse and the process of Centering. *Computational Linguistics*. 20(2):193-2