Ellipsis resolution for disguised agent

Shigeko Nariyama
University of Melbourne
/ Nara Institute of Science and Technology
Information Science, 8916-5 Takayama
Ikoma, Nara, Japan 630-0101
shigeko@unimelb.edu.au

Abstract

There is a type of nominal ellipsis that has been neglected in the study of ellipsis resolution. Certain sentence constructions, observed in diverse languages, tend to disguise the existence of ellipsis - due to an anomalous case assignment that assigns the nominative case to what is not typically the subject. This phenomenon deserves attention, for the referent of such an ellipsis is the agent of the sentence that carries vital information for Question Answering. Furthermore, as many languages including English do not use these constructions, resolution of such ellipses is of importance for Machine Translation and other multilingual applications. This paper presents a method to detect such disguised ellipses and to resolve them as well as common ellipses in a unified manner.

1 Introduction

The study of nominal ellipsis resolution differs significantly from that of pronominal and other anaphoric resolution with the view that ellipsis in its own right has null morphology, and hence the detection of ellipsis is an additional task. Consequently, much work on nominal ellipsis resolution has focused on resolving syntactically missing arguments, i.e. implicit subjects, and to a much lesser extent, objects and indirect objects. Since the subject is the most frequent and usually the most semantically significant argument, this focus is legitimate. Even so, current research is still challenged by the prospect of adequately accounting for even subject ellipsis. Needless to say, there have been attempts to resolve other types of nominal ellipsis, such as lexical nouns, indirect anaphors, and exophora (Mitkov 2002).

This paper is an elaboration of the paper (Nariyama 2003a) that raised another type of ellipsis whose resolution has been hitherto neglected, presumably because the existence of this type of ellipsis is even more inconspicuous than those ellipses that have been studied. It occurs in some constructions, which appear to be syntactically complete but are in actuality missing an agent. Hence, this disguised ellipsis is referred to as 'incognito' ellipsis. As the agent is the key nominal constituent that is informationally vital, particularly in applications for Question Answering and Machine Translation systems, it deserves attention and sound analysis.

The rest of this paper is organised as follows. Section 2 lays out background of the phenomenon. Section 3 examines the aspect of detecting such ellipses. For acheiving this task, it discusses the process of creating a list of argument taking nominals in order to alleviate the problems arising from word sense disambiguations inherent in the semantics of verbs. Finally, Section 4 presents a method for resolving incognito ellipses as well as common ellipses in a unified manner.

2 Incognito ellipsis

The constructions that contain 'incognito' ellipsis are observed in diverse languages, such as Spanish, Russian, Latin, Japanese, Turkish, and Old English (Shibatani, 1982:106). The following examples from Japanese newspaper articles demonstrate the point.[1] Both (1) and (2) appear syntactically complete, containing all subcategorised arguments, including, and most importantly, the subject of the sentence. However, semantically there is a missing agent. Indeed, in languages, such as English, an overt expression of the agent is required, as demonstrated by the fact that the literal translation of these sentences produces peculiar English sentences. The natural English translation (shown by the arrows →) requires a specification of the agent that is not expressed in the Japanese sentences; namely, 'who has the intention' in (1) and 'who does not understand' in (2). In other words, despite its nominative marking that prototypically denotes the subject, the nominatively marked nominal is essentially the object of the sentence.

[1] The examples only show the part of the sentences that succinctly demonstrates the points of argument, instead of quoting the whole long sentences.

(1) *Bosnia–o hoomonsuru ishi-ga aru.*
Bosnia-OB visiting intention-SB exist/be

(lit.) 'An intention to visit Bosnia exists. / There is an intention to visit Bosnia.'
→ 'ø has the intention to visit Bosnia.'

(2) *Shin'i-ga wakara-nai.*
true intention-SB understand-not
(lit.) 'The real intention is not understandable.'
→ 'ø don't understand the real intention of ø.'

Analogously, Spanish expresses (1) and (2) as (3) and (4) respectively:

(3) *Hay intención de visitar Bosnia.*
Exist intention of visit Bosnia
(lit.) 'Intention of visiting Bosnia exists'.

(4) *No se sabe cuál es la verdadera intención*
Not self know what is the true intention
(lit.) 'The true intention is not known itself.'

Subject ellipsis is prevalent in Spanish and many other European languages, and ellipsis resolution is achieved commonly by verbal inflection reflecting number, person and gender of subject ellipsis. However, the verb *hay* in (3) has no agreement with the syntactic subject, *intención*. Moreover, analogous to Japanese examples, *intención* functions more like the object. Indeed, it takes an accusative pronoun when pronominalised. The same applies to (4).

This type of construction and ellipsis is prevalent, at least in Japanese (12% of clauses are of this type, and 87% of those incognito agents are implicit; see Table 1 in Section 2.2).

As demonstrated above, a problem is posed by the fact that many languages, including English, do not utilse those constructions, and require overt expression of the referent that is incognito in those languages which use 'incognito ellipsis' constructions. Hence, the recognition and resolution of such ellipses is of importance particularly in multilingual applications including Machine Translation, and in Question Answering and Information Retrieval for capturing the agent.

This paper focuses on the phenomenon seen in Japanese. There have been numerous papers on resolving Japanese ellipsis (Nakaiwa et al. 1995; Walker et al. 1994; Kameyama, 1985; inter alia). A book by Mazuka and Nagai (eds.) (1995) collects 14 papers addressing different aspects of Japanese sentences from the point of view of sentence processing. However, the phenomenon of 'incognito ellipsis' raised in this paper is not dealt with in those papers.

2.1 Agent disguising constructions

Some constructions disguise an agent even in English. This is summarised as follows:

I. Intransitive constructions (e.g. 5b)
II. Passive constructions (e.g. 6b)
III. Constructions with pleonastic pronoun (e.g. 7)
IV. Existential constructions (e.g. 1,3,8,9)
V. Nominative object constructions (e.g. 2,4)

The following pairs of sentences show that transitive sentences can be alternated with intransitive and passive sentences, in effect reducing the valency of the sentence, i.e. in (5b) and (6b) the agent is not expressed:

(5a) I opened the door.
(5b) The door opened.

(6a) I held a meeting yesterday.
(6b) A meeting was held yesterday (by me).

These types of constructions and their relationship with the corresponding transitive sentences that are also observed in English are well documented; hence this is not the problem that this paper is addressing. Similarly, in the third constructions (see 7), the agent is non-specific, generic or obfuscated and the implicit agent of 'resolve' is not generally treated as elliptical.

(7) It is difficult to resolve pronominal anaphors.

The problem is posed by the next two types of constructions. The fourth type is the existential constructions, "There is/are … ." in English. In Japanese and many other languages, this type of construction is also used to express one's possession, in which case the possessor needs be identified when implicit, for example:

(8) Ø *Imooto-ga iru.*
sister-SB exist/be
(lit.) 'A/my sister exists. / There is a sister.'
→ 'ø have a sister.'

(8) appears complete syntactically, as it has the subject in the intransitive sentence. However, analogous to (1), semantically there is a missing possessor. As can be seen, the literal English translation of the Japanese sentence, 'A/my sister exists' or 'There is a sister', may be grammatical but strange or carry another meaning. English expresses the proposition by the 'have' construction, i.e. 'I have a sister.'

Discourse structure has a bearing on the differences. The existential constructions in

English provide background information into a discourse by introducing a new indefinite referent (Insua 2003). Thus, anaphoric referents tend not to be expressed by those constructions. However, this is not the case in Japanese; the existential constructions express anaphoric referents, which are often implicit. In other words, the same sentence construction may convey different anaphoric information in different languages.

Furthermore, (8) can express the physical presence of a 'sister' as in English. This can be distinguished to a greater extent by the presence of locational or temporal adverbs, as in 'My sister is over there now,' in which case it is a prototypical existential sentence and there is no ellipsis, conceding that nonetheless the precise distinction between the two meanings must be investigated more precisely.

The fifth is the nominative object constructions, of the kind that are not observed in modern English. In these constructions, the prototypical subject is marked by the dative (*–ni*, (or *–de, -ga*) in Japanese) or the topic marker (*-wa*), while the prototypical object is marked by the nominative (*-ga*), which predominantly marks the subject, hence the name 'nominative object' (NomOB). That is, having a construction of: [X-*wa/ga/ni(wa)/de(wa)* Y-*ga* Verbal] (Verbal: verbs, adjectives, or nominal adjectives).

For both the existential constructions and the nominative object constructions, the problem arises from the fact that X is often implicit and only [Y-$_{SB}$*ga* Verbal] is overtly expressed, missing an agent. Hence, syntactically it is difficult to detect the existence of the agent.

Things are further complicated, in that some constructions can involve two ellipses (see §3.2):

(9) *Denwa-ga atta.*
telephone-SB existed
(lit.) 'There was a call (for X from Z).'
➔ '(Z) called (X).'

Figure 1 summarises the two types of argument structures that contain incognito ellipsis – Single (1, 2, 8) and Double (9). Section 3 examines the structures from various aspects and ways to detect incognito ellipsis for machine processing.

Single incognito ellipsis construction (NomOB(+1)):
[(X_{Agent}) Y-$_{SB}$*ga* Verbal]

Double incognito ellipses construction (NomOB(+2)):
[(Z_{Agent}) ($X_{Recipient}$) Y-$_{SB}$*ga* $Verb_{intoransitive}$]

Figure 1: Argument structures with incognito ellipsis

2.2 Implications of incognito ellipsis

As a reference point for getting some idea of the extent and prevalence of incognito ellipsis, Nariyama (2003a) reported the results (Table 1) from examining 5 newspaper articles and 5 magazine articles (PHP).

	newsp	PHP	total
# of clauses/sentences	110/36	410/172	520/208
# of ellipsis	66	267	333
# of ellipsis/sentence	**1.83**	**1.55**	**1.60** ①
# of incognito ø / incognito agent	9 / 11	46 / 52	55/63
% of incognito ø	**81.8%**	**88.5%**	**87.3%** ③
% of incognito ø / all ø	**13.6%**	**17.2%**	**16.5%** ②

Table 1: Proportion of incognito ellipsis and related figures

Table 1 shows that in total 333 ellipses are found in 520 clauses; ① on average every sentence contains 1.6 ellipsis; ② 16.5% of those ellipses are incognito ellipses; and ③ the incognito agents are ellipted on average 87.3%. In other words, 12.1% (63/520) of clauses form the constructions that involve incognito agents; 87.3% of these are ellipted; and 16.5% of all ellipses derive from those constructions.

As a further reference, (2) was translated into English using two machine translation systems, to ascertain how sentences under the constructions are currently translated. As resolution of ellipsis requires contextual information, the referent and context are added to (2) to make (10). The translated sentences reveal that 'real intention' is treated as the subject of the sentence, and the incognito agent is not resolved.

(10) *Watashi-wa shachoo-no hanashi-o kiita ga, ø shin'i-ga wakara-nai.*
'I listened to the president's talk, but (I) don't understand (his) true intention.'

MT1: ? 'I heard the story of the president, but real intention is not understood.'

MT2: ? 'Real intention is not understood although I heard the president's talk.'

3 Detecting incognito ellipsis constructions

3.1 Single incognito ellipsis construction

Single incognito ellipsis construction is laid out in Figure 2.

[(X-*wa/ga/ni(wa)/de(wa)*) Y-$_{SB}$*ga* Verbal]
-Top/Nom/Dat(Loc)/Ins/ -Nom
(Agent) (NomOB)

Figure 2: Single incognito ellipsis construction

'Verbal' generally denotes possession, as in (1) and (8), ability (11), state (12), and volition.

(11) *(Nihon-wa) anzenhoshoo mende kokuren-ni*
Japan-Top security area the U.N.-to
Nanraka-no kooken-ga dekiru.
Something-of contributions-SB can

(lit.) '(As for Japan,) some contribution to the U.N. in the area of security is possible.'
→ '(Japan) is able to contribute in some way to the U.N. in the area of security.'

(12) *(John-wa) seikaku-ga warui.*
(John-Top) personality-SB bad
(lit.) '(As for John, his) personality is bad.'
→ '(John) has a bad personality.'

(12a) *John-ga warui.*
John-SB bad
'John is bad.'

It is cutomery to subcategorise arguments from the semantics of verbals. However, it is apperant from (12) and (12a) that the semantics of the nominative object as well as the verbal must be accounted for to capture constructions with incognito ellipsis. (12a) has the same verbal and identical surface structure to (12), but (12a) does not contain ellipsis.

Note that some nominative objects, as in (2), require a possessor, which is also implicit, since marking of the determiner is not grammatically required in Japanese (see the studies on (*A no*) *B* 'B (of A)' and 'a/the B' by Bond 2001, Murata and Nagao 1999). Sentences such as (2) are included in Single incognito ellipsis construction despite having two incognito ellipses, because of their construction type that belongs to Single incognito ellipsis construction rather than Double.

(2) *(X-wa) (A-no) shin'i-ga wakara-nai.*
-of true intention-SB understand-not
'X don't understand the real intention of A.'

Given these complications despite the linguistic cues, it is formidably difficult to automatically and reliably detect implicit agents and possessors from the semantics alone, and to differentiate those sentences that involve incognito ellipsis from those that do not (such as proto-typical intransitive/ existential sentences). So a list of argument taking nominals in addition to verbal subcategorisation will help to detect and resolve incognito ellipsis, and alleviate the problems arising from word polysemy (hence the need for word sense disambiguations) inherent in the semantics of verbals.

A study is underway to examine approximately 6000 verbals in *Goi-taikei* (Japanese lexicon dictionary, Ikehara et al. 1997), using the linguistic cues to single out lexica that involve incognito arguments, and then manually reexamine the lexica. This will be followed by the examination of nominals for the semantic attributes of nominative objects.

3.2 Double incognito ellipsis construction

Double incognito ellipsis construction forms the structure, $[(Z_{Agent})\ (X_{Recipient})\ Y\text{-}_{SB}ga\ Verb_{intransitive}]$, as seen in (9). The occurrence of this structure is not limited with the existential verbs, for instance:

(13) *Houkoku-ga todoita/kita/haitta.*
report-SB arrived/came/entered
'A report arrived/came/came in (for X) (from Z).'

Analogous to Single incognito ellipsis construction, whether or not a verb subcategorises for two participants X and Z as well as nominative object Y depends essentially upon the semantics of not only the verb but also Y. So, for example, *Kuruma-ga kita* 'A car came' is an intransitive sentence without X and Z (save location), and its only difference with (13) is the type of Y. Thus, another list of argument taking nominals needs to be created in order to alleviate the problems of word sense disambiguations.

3.2.1 Creating a list of argument taking nominals

This subsection categorises the types of verb and nominative object Y, which saves work from examining all verbs and nouns in creating this list.

First, apart from the existential verbs, the motion verbs are typically of this type that subcategorise two arguments. So the motion verbs are examined using *Goi-taikei* thesauraus. The interim report on the examination suggests that there are only handful of motion verbs that takes two participants X and Z. They are: *kuru* 'to come', *todoku* 'to arrive', *oriru* 'to be granted', and *hairu* 'to enter'. These verbs are centripetal, having a deictic center and the movement is directed towards the center. Other motion verbs are not of this type for two reasons. One is that they have a movement but do not have the deictic center; for example, *tobu* 'to fly' and *susumu* 'to proceed'. The other is centrifugal, *iku* 'to go' and *wataru* 'to cross', whose movement goes away from the center, which requires a different treatment from the centripetal verbs.

Secondly, those nominative objects Y that subcategorise X and Z are commonly communications related words, such as *renraku* 'contact' and *tegami* 'letter' and benefactive, such as *okurimono* 'gift' and *oiwai* 'celebratory gift'. Lexically they are generally Chinese originated verbal nouns (*sahen meishi*), such as *denwa* 'phone call', deriving from 'to call' and *henji* 'reply' deriving from 'to reply', and verbal nouns, such as *sasoi* 'invitation' and *moshide* 'offer'. These Y involve two people or organisations that represent people in it. So, for example, 'offer' must involve two people: one who makes an offer and the other who receives the offer.

Goi-taikei thesauraus and 4158 verbal nouns are being examined for making a list of two-argument taking nominals. In addition, sentences collected from a newspaper corpus are being examined for their collocations. The difficulty of this analysis is due to the fact that only overt nominals are listed, i.e. ellipsis is not indicated. So having no contexts in the corpus, this approach takes a lot of inferencing.

Given the facts that the verb imposes selectional restrictions on the type of Y (i.e. each verb selects different Y, although *kuru* 'to come' and *aru* 'to be' can be combined with most Y) and the Y in turn does the same to X and Z, and that there are only handful of the verbs, it pays to make a list of argument structures in the form of [Y Verb] as a set consitituent with specified lexica. This will further alleviate the problems of word sense disambiguations.

Owing to those semantic restrictions, Double incognito ellipsis construction tends to appear frequently on the topic of communications and gift giving. For example, the news about the Japanese hostages in Iraq (April 2004) innumerably used the constructions reporting the communications among the people detaining the hostages, their religious leader, the Japanese embassy in Iraq and the Japanese government in Japan.

3.2.2 Linguistic mechanisms for differentiating two arguments

Resolving the two arguments X and Z is problematic, in that since the semantic attributes of both X and Z are identical, typically referring to humans and the like, ellipsis resolution of this type is unwieldy under the systems relying solely on selectional restrictions, and may be unreliable with stochastic models alone. Hence, more detailed linguistic mechanisms behind the constructions that bear incognito ellipsis should be investigated for better ellipsis resolution.

Why is it that those constructions are used instead of transitive sentences, namely, '(Z) called (X)' for (9), and '(Z) sent a report (to X)' for (13)? Nariyama (2003b) identifies the reasons to be the organisation of discourse structure. She has proposed a set of principles applicable to Japanese that captures the pattern of ellipsis (Principle of ellipsis), which in turn is governed by the structure of sentences (Principle of direct alignment).

Japanese sentences are structured in such a way as to express an argument high in animacy [1>2>3>animate>inanimate] and discourse salience (topicality/prior mention) as the subject. In other words, a subject must be higher than non-subject arguments (SB>nonSB) in terms of animacy and discourse salience. So for example, 'I called my husband', which has a first person subject and a third person object, i.e. [1→3] forming a direct alignment, is acceptable. On the other hand, the reverse, 'My husband called me', ([3→1]) violates the principle of direct alignment, and indeed the sentence sounds unusual and infelicitous in Japanese, although it is acceptable in English. Instead, the proposition is often expressed using the existential construction, 'There was a call for me from my husband' as in (9); i.e., using Double incognito ellipses construction. Conversely, the existential construction cannot be used for sentences with direct alignment, i.e. 'There was a call for my husband from me.' In other words, the construction type signals the relationship between X and Z. The use of the existential (i.e. intransitive) construction signals that, in the case of (9), the caller Z is lower than the receiver X of the call in terms of animacy and discourse salience, while a transitive sentence signals that the caller is higher than the receiver.

This restructuring process is formulated in Figure 3, which is built into the makeup of the algorithm presented in Section 4.1.

$[(Z_{SB})\ (X_{IO})\ Y_{OB}\ Verb_{transitive}]$
where Z>X in terms of animacy and discouse salience

$[(X_{for})\ (Z_{from})\ Y_{SB}\ Verb_{intransitive}]$
where Z<X in terms of animacy and discouse salience

Figure 3: Restructuring formula

Based on the principles, four ellipses in (14) are easily resolved without resort to commonsense knowledge or the cause-effect logic. The subordinate clause is a transitive sentence, so the agent is higher than the recepient in terms of person or discourse salience, while the matrix clause is intransitive so the agent is lower. This leads to the correct reading indicated.

(14) *Renrakushita ra, henji-ga kita.*
contacted when reply-SB came
'When (I) contacted (him), (he) sent (me) reply.'

The claim that sentences that violate the principle of direct alignment are restructured into the existential or nominative object constructions (i.e. intransitive sentences) is verified in the small bilingual corpus analysis; intransitive sentences are used 42.8% in Japanese, while 23.4% in the English translations (Nariyama 2003a).

Moreover, the principle of ellipsis specifies that the higher an argument in terms of animacy and discourse salience, the more prone it will be to ellipsis. Indeed, when a higher argument is ellipted, as in 'There was a call (for me) from my husband', the sentence sounds natural. On the other hand, when a lower argument is ellipted while a higher is retained, as in 'There was a call for me (from my husband)', the sentence is unnatural in Japanese, sounding as if there should have been an argument specified. Thus, the overt argument works as a reference to resolve ellipsis.

4 Resolution of ellipsis

4.1 Algorithm

This paper adopts one simple linguistically oriented method and algorithm for ellipsis resolution demonstrated in Nariyama (2003b), which is said to resolve the bulk of nominal ellipsis. For simplicity given the scope of this paper, only the part of the algorithm pertinent to the topic addressed in this paper is presented. Owing to lack of corpus that are tagged with ellipses including incognito ellipses, this paper presents the method and reports the results of a hand-tested newspaper articles and magazine articles reported in Nariyama (2003a).

Because language use and its mechanism are not only formidably complicated but also vagarious, once the lists of verbals and argument taking nominals that involve incognito ellipsis have been created, the proposed method will then be auto-tested, and combined with a learning and stochastic model, such as the tournament model by Iida et al. (2003) for improving performance.

The core of the algorithm is captured in the use of "salient referent list". This is like a memory bank that pools old referents from the previous sentences, attempting to reflect how humans may store referential information, and hence it builds context and inference. It is this input information that provides cues to resolve various types of ellipsis, including subject ellipsis, non-subject ellipsis and multiple ellipses (more than one ellipsis per clause) for inter as well as intra-sentencial ellipsis. This paper extends its application to resolving incognito ellipsis.

Salient referent list details all overt arguments in the sentence, merged with the arguments that have appeared up until the sentence in question. These arguments are listed following "salient referent order list",[2] which accords TopSB (topicalised subject) the highest saliency. NomOB (nominative object) is included in the list, in order to account for those constructions with an incognito agent.

TopSB (Global > Local > Quotation) > TopOB > SB > IO > NomOB > OB > Other

Figure 4: Salient referent order list

4.2 Resolving ellipsis

This subsection explains the way in which salient referent lists (SRL) are created and used to resolve ellipses using fragments of a newspaper article. Each clause is indicated by square brackets [ ] with the clause number on the right side. The matrix clause is numbered but not bracketed.

[s1]
Mr K_a *-wa* *Mr* S_b *to* *kaidanshita.*$_1$
-TopSB with had a meeting
'Mr K_a had a meeting with Mr S_b.'

[s1] has two overt arguments –Mr K (TopSB) and Mr S (Other). By following the salient referent order list, TopSB is listed higher than Other. Each listed argument is given a number. The argument under T1 has the highest saliency and is therefore the best candidate as referent for the ellipsis; T2 is the next highest, and so forth. These arguments are listed in SRL accordingly, provided with detailed lexical information, including the grammatical relation, topicality, animacy (and semantic attributes using *Goi-taikei*, the detail of which is not shown below). Accordingly, the SRL for [s1] is formulated:

SRL [s1] {$T1_a$: Mr K (TopSB; 3rd person) >
$T2_b$: Mr S (Other; 3rd person)}

Ellipsis is resolved based on the information contained in the SRL for the whole sentence, instead of a clause, where the ellipsis appears. [s1] contains no ellipsis, so we process the next sentence [s2].

[2] The salient referent list and its order were eclectically adapted from variety of sources, including the Japanese version of Expected Center Order in Centeing Theory (Kameyama, 1985), and Givon's Topicality hierarchy (1979). The idea was in turn rooted in the hierarchy proposed by Keenan and Comrie (1977). This paper's use of such a linguistically fundamental hierarchy made some resemblance to other approaches, for example, Lappin and Leass (1994).

[s2]
[ϕ_a ϕ_c shin'i$_c$-ga wakara-nai]$_{1/LT1}$
SB Poss true intention-NomOB understand-not
to ϕ_a shinchoona shisei$_d$-o miseta.$_2$
that SB cautious attitude-OB showed
→ '(He$_a$) appeared cautious by saying that (he$_a$) didn't understand (his$_b$) real intention.'

SRL is updated with each new sentence. [s2] has two overt arguments: *shin'i* (NomOB(+1), which is detected from the semantics of NomOB and the verb *wakaru* 'understand' that requires an agent of 'understand', and noted as (+1) for having single ellipsis, not double) and *shisei* (OB). These are incorporated into the SRL [s1] to make SRL [s2] by following the salient referent order list.

SRL [s2] {T1$_a$: Mr K (TopSB; 3rd person) >
T2$_c$: *shin'i* (NomOB(+1); inanimate) >
T3$_d$: *shisei* (OB; inanimate) >
T4$_b$: Mr S (Other; 3rd person)}

The semantics and the structure of the clauses signal that [s2] has three ellipses: the subject and the possessive in Clause 1 and the subject in Clause 2. Ellipsis is resolved per clause. So the T1 argument is applied as referent to each clause. Clause 1 contains two ellipses, which are also ordered following the salient referent order list. As the subject is higher than the possessive (Other), T1 is applied to the subject and T2, the next highest, is applied to the other ellipsis. However, T2 is incompatible because the possessor cannot be the same as the possessed. So T3, the next down, is examined. The semantics of T3 is incompatible with those selected by the head noun *shin'i*. So T4, the next down, is chosen. This interpretation, following the method, selects the correct referents for ellipses including incognito ellipsis.

This operation reflects the mechanisms pr esent-ed in §3.2.2 that the argument high in animacy and discourse salience is expressed as the subject (which is listed high in SRL) and most prone to ellipsis (i.e. taking the highest argument in SRL as referent).

[s3]
Futuka go, Mr S$_b$ kara ϕ_a renraku$_e$-ga atta.$_{1/LT2}$
2 days later, from contact-NomSB existed
(lit.) 'Two days later, there was a contact (for him$_a$) from Mr S$_b$.'
→ 'Two days later, Mr S$_b$ contacted (him$_a$).'

[s3] contains two overt arguments: Mr S (SB) and *renraku* 'contact' (NomOB), which replaces the NomOB in SRL [s2] for the reasons of recency, as the current system lists only one argument under any one grammatical relation. Note that Mr S is listed as SB, because, as shown in Figure 3, Z in 'from Z' of NomOB(+2) is in essense the subject. Accordingly, SRL [s3] is created as follows:

SRL [s3] {T1$_a$: Mr K (TopSB; 3rd person) >
T2$_b$; Mr S (SB, 3rd person)
T3$_e$: *renraku* (NomOB(+2); inanimate)>
T4$_d$: *shisei* (OB; inanimate)}

As the semantics of the NomOB 'contact' select two arguments (source and goal), [s3] involves double incognito ellipses, noted as 'NomOB(+2)' in the SRL. So [s3] is missing a goal argument. The T1 argument is chosen as the referent, which provides the correct interpretation.

4.3 Test results and evaluation

SRL was hand-tested on the same texts used for Table 1. The results are shown in Table 2.

	newsp	PHP	total
# of clauses/sentences	110/36	410/172	520/208
# of ellipsis	66	267	333
# of ellipsis/sentence	1.83	1.55	1.60
# of incognito ø / incognito agent	9 / 11	46 / 52	55/63
% of incognito ø	81.8%	88.5%	87.3%
% of incognito ø / all ø	13.6%	17.2%	16.5%
④SRL: % of ✓ for all ø	**87.9%**	**82.0%**	**85.6%**
SRL: % of ✓ for incognito ø	100%	71.7%	77.2%

Table 2: Result of Salient Reference List

④ in Table 2 shows that the use of SRL resolves ellipses, including incognito ellipses, with an accuracy of 85.6%. 48 incorrect selections were made by SRL, which were due to the following reasons. The most frequent mistakes were caused by the method not distinguishing generic referents from particular referents.

1. 22/48 (6.6% of all ellipses): ellipsis referring to generic referents that do not appear in the context. All of these occurred in PHP.
2. 13/48 (3.9%): number problem. SRL selects "I", but from the context "we" is appropriate.
3. 11/48 (3.3%): two topics simultaneously dominating the story. They are distinguished by contextual and commonsense knowledge.
4. 2/48 (.6%): referring to the whole sentence, instead of a particular referent.

4.4 Other studies on ellipsis resolution

Since the phenomenon of incognito ellipsis has not been raised in the study ellipsis resolution, the algorithm presented in §4.2 can be fully compared with no methods. So the overall method is

compared here with that of Centering theory by Walker et al. (1994), as Centering theory is a widely adopted theory and also has a focus-based approach.

Centering theory seems to be in essence a theory that explains the interpretation of an anaphor for having a particular referent as its antecedent. It is also used to predict a perferred reading based on four transitional states: Continue, Retain, Smooth-shift, and Rough-shift. With the assumption that sentences maintain discourse coherence, the theory selects a preferred reading that has the least shift (i.e. Continue) of topic (attentional state) to be the correct reading, and introduced a constraint 'Zero topic assingment' to maintain Continue. This claim was substanciated in the same paper using constructed sentences. However, Iida (1998), one of the authors of Walker et al. (1994), examined newspaper articles and found that a shift of topic is abundant even with the use of zero (ellipsis) (60/136 shifted). Thus the result refutes the claim of Walker et al. (1994).

Other shortcomings as I see in Centering theory for the application of ellipsis resolution include lack of clarity and adequacy in handling complex sentences and multiple ellipses. The method to retain one and only one Cb from the immediately preceding sentence is also problematic, which is also problematic for the proposed method of this paper. The solution to this problem requires inclusion of global topic and some contraints that regulate a switch of interpretation from local topic to global, as well as the recognition of exophora.

5 Conclusion

This paper raised the need for resolving a neglected type of nominal ellipsis, 'incognito' ellipsis, whose constructions are apparently syntactically saturated but which nevertheless require resolution for full interpretation.

Lists of verbals and nominals that subcategoirse for one or two incognito ellipses are being created. These lists will lessen the difficulty of capturing incognito ellipsis differentiating from optional deletion (in proto-typical intransitive sentences) for surface-structurally identical sentences for machine processing. Given the complexity and vagary of human language, the proposed method will be combined with a learning/stochastic model and be tested it in large corpus.

6 Acknowledgements

I wish to thank Tim Baldwin, Ryu Iida, Connal Parsley, Mikel Forcada and anonymous reviewers for their inputs. This research report was supported by JSPS 02348.

References

Francis Bond. 2001. *Determiners and number in English contrasted with Japanese, as exemplified in machine translation*, Ph.D. thesis, University of Queensland.

Talmy Givón. 1979. *On understanding grammar*, Academic Press, New York.

Masayo Iida, 1998. Discourse coherence and shifting centers in Japanese texts. In Walker et al. *Centering theory in discourse.* pages 161-180.

Ryu Iida, et al. 2003. Incorporating contextual cues in trainable models for coreference resolution, EACL Workshop on *Computational treatment of anaphora*, Budapest, pages 23-30.

Satoru Ikehara et al. 1997. *Goi-Taikei – Japanese Lexicon*. Iwanami Publishing, Tokyo

Ana Insua. 2003. Present day English existential *there*-constructions and their pragmatics towards an integrated categorization, ms.

Megumi Kameyama. 1985. *Zero anaphora: the case of Japanese*, Ph.D. thesis, Stanford University.

Edward Keenan and Bernard Comrie. 1977. Noun accessibility and universal grammar, *Linguistic Inquiry* 8:63-99.

Shalom Lappin and Herber J. Leass. 1994. An algorithm for pronominal anaphora resolution, *Computational Linguistics* 20(4):535-561.

Mazuka R. and Nagai N. (eds.). 1995. *Japanese sentence processing*, Lawrence Erlbaum, NJ.

Ruslan Mitkov. 2002. *Anaphora resolution*, Longman, London.

Masaki Murata and Makoto Nagao. 1999. Resolution of indirect anaphora in Japanese sentences using examples 'X no Y' (Y of X), ACL-99 Workshop on *Coreference and its applications*, pages 31-38.

Hiromi Nakaiwa. et al. 1995. Extrasentential resolution of Japanese zero pronouns using semantic and pragmatic constraints, *AAAI-95 Symposium*, pages 99-105.

Shigeko Nariyama. 2003a. Resolving incognito ellipsis: treatment for constructions that disguise ellipsis, In Proceedings of *Machine Translation Summit IX*. New Orleans, pages 261-268.

Shigeko Nariyama. 2003b. *Ellipsis and reference tracking in Japanese*, SLCS 66, John Benjamins.

Masayoshi Shibatani. 1982. Japanese grammar and universal grammar, *Lingua* 57:103-123.

Marilyn Walker. M. Iida and S. Cote. 1994. Japanese discourse and the process of Centering, *Computational Linguistics* 20(2):193-232.