Determination of referential property and number of nouns in Japanese sentences for machine translation into English

Masaki Murata, Makoto Nagao

Department of Electrical Engineering,
Kyoto University, Kyoto, JAPAN
{mmurata,nagao}@kuee.kyoto-u.ac.jp

Abstract

When translating Japanese nouns into English, we face the problem of articles and numbers which the Japanese language does not have, but which are necessary for the English composition. To solve this difficult problem we classified the referential property and the number of nouns into three types respectively. This paper shows that the referential property and the number of nouns in a sentence can be estimated fairly reliably by the words in the sentence. Many rules for the estimation were written in forms similar to rewriting rules in expert systems. We obtained the correct recognition scores of 85.5% and 89.0% in the estimation of the referential property and the number respectively for the sentences which were used for the construction of our rules. We tested these rules for some other texts, and obtained the scores of 68.9% and 85.6% respectively.

1   Introduction

One of the difficult problems in machine translation from Japanese to English or other European languages is the treatment of articles and numbers. There are referential pronominals in Japanese such as KONO, ANO, etc., but these are used only in particular occasions where references are to be indicated definitely. As to the number the Japanese language has no plural form for nouns and no distinction in verb conjugation to indicate the number of subject or object of a verb. In English there are definite and indefinite articles for nouns and also the distinction between singular and plural. Therefore the correspondence of articles and numbers for nouns in Japanese to English translation is a very difficult problem.

To solve this problem to a certain extent, we have to estimate the referential properties of nouns in a sentential utterance. It is commonly believed that the language understanding mechanism is necessary to solve this problem, and certain contextual or inter-sentential information is to be grasped. It is true, but it is difficult at the present level of natural language analysis technology.

We propose here that lots of keys exist in the surface information of a sentence to determine the referential property and the number of a noun in the sentence. For example, "KARE-WA(he) GAKU-SEI(student) DESU(is)" indicates that KARE is a specific person(singular), and is linked by a copula to GAKUSEI, which is a countable noun. Therefore the property, singular, is inherited to GAKUSEI from KARE, and the translation is "He is a student". When the above example is changed as "KARE-WA(he)

KINOU(yesterday) ITTO(first) SHOU-O(prize) MORATTA(was given) GAKUSEI(student) DESU(is)", where "student" is modified by an embedded sentence "he was given the first prize yesterday" , this indicates that "student" in this sentence is strictly specified, and is definite. Therefore the English expression to this Japanese expression is "He is <u>the</u> student who was given the first prize yesterday".

This sort of judgement is not absolutely reliable but just probable. This means that what we have to do is to construct a kind of expert system by incorporating large number of heuristic rules with certain reliable factors. In the following we will describe what kind of heuristic rules we have written for the articulation of the referential property and the number of a noun in a Japanese sentence.

2 Categories of Referential Property and Number

2.1 Categories of Referential Property

Referential property of a noun phrase here means how the noun phrase denotes the subject. We classified noun phrases into the following three types from the referential property.

noun phrase
- **generic** noun phrase
- **non generic** noun phrase
  - **definite** noun phrase
  - **indefinite** noun phrase
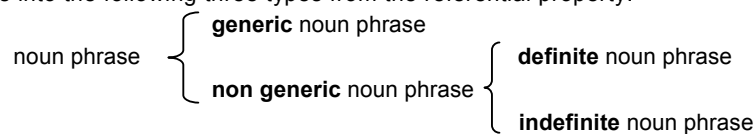
A noun phrase is classified as generic when it denotes all members of the class of the noun phrase or the class itself of the noun phrase. For example, "dogs" in the following sentence is a generic noun phrase.

<u>Dogs</u> are useful. (1)

A noun phrase is classified as definite when it denotes a contextually non-ambiguous member of the class of the noun phrase. For example, "the dog" in the following sentence is a definite noun phrase.

<u>The dog</u> went away. (2)

An indefinite noun phrase denotes an arbitrary member of the class of the noun phrase. For example, the following "dogs" is an indefinite noun phrase.

There are three <u>dogs</u>. (3)

2.2 Categories of Number

Number of a noun phrase is the number of the subject denoted by the noun phrase. Categories of number are as follows.

noun phrase
- **countable** noun phrase
  - **singular** noun phrase
  - **plural** noun phrase
- **uncountable** noun phrase
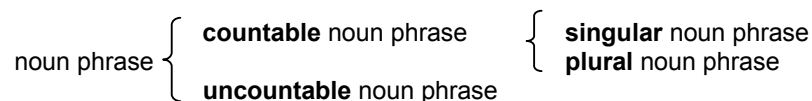
3 How to Determinate Referential Property and Number

Heuristic rules for the referential property are given in the form:
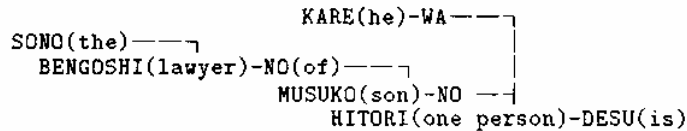*(condition for rule application)*
$\Rightarrow$ { indefinite*(possibility, value)* definite*(possibility, value)* generic*(possibility, value) }*
Heuristic rules for the number are given in the form:

```
"KARE(he)-WA SONO(the)-BENGOSHI(lawyer)-NO(of) MUSUKO(son)-NO(of)
HITORI(one person)-DESU(is)."      (He is one of the sons of the lawyer.)

                          (a):Japanese sentence

                                      KARE(he)-WA——┐
              SONO(the)——┐                         |
              BENGOSHI(lawyer)-NO(of)——┐           |
                               MUSUKO(son)-NO —┤
                                    HITORI(one person)-DESU(is)

              (b):Dependency structure of sentence(a)


( <[noun common-noun _ _ 'HITORI' 'HITORI']
   [copula _ copula DESU-line-basic-form 'DA' 'DESU']
   [punctuation-mark period _ _ '. ' '. ']>
   ( <[noun common-noun _ _ 'MUSUKO' 'MUSUKO']
       [postpositional-particle noun-connection-postpositional-particle _ _ 'NO' 'NO']>
       ( <[noun common-noun _ _ 'BENGOSHI' 'BENGOSHI']
           [postpositional-particle noun-connection-postpositional-particle _ _ 'NO' 'NO']>
           ( <[referential-pronominal _ _ _ 'SONO' 'SONO']> )))
   ( <[noun common-noun _ _ 'KARE' 'KARE']
       [postpositional-particle topic-marking-postposition _ _ 'WA' 'WA']
       [punctuation-mark komma _ _ ', ' ', ']> ))

              (c):Dependency structure representation of sentence(a)

          Figure 1: Example of dependency structure representation
```

```
( <[noun -] - >
   ( <[referential-pronominal _ _ _ 'SONO' 'SONO']> ) - )

        Figure 2: An expression of the noun modified by "SONO(the)"
```

(condition for rule application)
    ⇒ { singular*(possibility, value)* plural*(possibility, value)* uncountable*(possibility, value)* }
In *condition for rule application,* a surface expression is written in the form like in Figure 2. *Possibility* has value 1 when the categories: indefinite, definite, generic, singular, plural or uncountable, are possible in the context checked by the condition. Otherwise the value is 0 for *possibility. Value* means that a relative possibility value between 1 and 10 (integer) is given according to the plausibility of the condition that the *possibility* is 1. Larger value means the plausibility is high.

    The rules are all heuristic so that the categories are not exclusive. In a certain conditional situation both indefinite and generic are possible, and also both singular and plural can co-exist. In these cases, however, the possibility values may be different.

    Several rules can be applicable to a specific noun in a sentence. In this case the possibility values are added for individual categories and the final decision of a category for a noun is done by the maximum possibility value. An example is given in Section 4.1.

    When determinating the referential property and the number of nouns, the condition part is matched not for a word sequence but for a dependency structure of a sentence. The dependency structure of

```
( <[noun common-noun _ _ 'HITORI' 'HITORI' indefinite singular]
   [be-verb _ be-verb DESU-line-basic-form 'DA' 'DESU']
   [punctuation-mark period _ _ '. ' '. ']>
   ( <[noun common-noun _ _ 'MUSUKO' 'MUSUKO' definite plural]
      [postpositional-particle noun-connection-postpositional-particle _ _ 'NO' 'NO']>
      ( <[noun common-noun _ _ 'BENGOSHI' 'BENGOSHI' definite singular]
         [postpositional-particle noun-connection-postpositional-particle _ _ 'NO' 'NO']>
         ( <[referential-pronominal _ _ _ 'SONO' 'SONO']> )))
   ( <[noun common-noun _ _ 'KARE' 'KARE' definite singular]
      [postpositional-particle sub-postpositional-particle _ _ 'WA' 'WA']
      [punctuation-mark komma _ _ ', ' ', ']> ))
```

Figure 3: The result of analyzing the sentence in Figure 1

a sentence (Figure 1(a)) is shown in Figure 1(b) which is represented as Figure 1(c)[1] to which the condition is checked. In heuristic rules, this expression can include a wild card(represented by "-") which can match any partial dependency structure representations. For example, a noun modified by "SONO(the)" is expressed as in Figure 2. There are many other expressions such as regular expressions, AND-, OR-, NOT-operators, MODee-operator for checking modifyer-modifyee relation and so on.

Algorithm of the Determination of a Category

The following steps are taken for the decision of a category for the referential property and the number.

(1) Sentences are transformed into dependency structure representations.

(2) Decision is made for each noun from left to right in the sentences transformed into dependency structure representation. This process allows the decision process to make use of the referential property and the number already determined (see 4.1(c)(d) for example). For each noun, the referential property is first determined, and then the number. This brings the utilization of referential property of a noun when analyzing the number of the noun (see 4.2(3) for example). In these processes all the applicable rules are used, *possibility* and *value* of each category are computed, and the category for the maximum value is obtained. An example of the result is shown in Figure 3. We can also utilize the global information of a document to which a sentence belongs in the decision process. The *condition* part, for example, can check whether there are identical nouns before. This information is useful for the determination of the referential property.

4 Heuristic Rules

We have written 86 heuristic rules for the referential property and 48 heuristic rules for the number. More than half of these rules are just the implementation of grammatical properties explained in standard grammar books of Japanese and English[2][3][4], but there are many other heuristic rules which we have originally introduced ourselves. Some of the rules are given below.

---

[1] This is the result transformed by the system[1].

4.1   Heuristic Rules for Referential Property

1. When a noun is modified by a referential pronoun, KONO(this), SONO(its), etc.,
   then   { indefinite (0, 0)[2]   definite (1, 2)   generic (0, 0) }
   Examples: KONO (This) HON-WA(book) OMOSHIROI(interesting)
   _____This book is interesting._

2. When a noun is accompanied by a particle (WA), and the predicate has past tense,
   then   { indefinite (1, 0)   definite (1, 3)   generic (1, 1) }
   Example: INU-WA(dog) MUKOUE(away there) IKIMASHITA(went)
   _____The dog went away._

3. When a noun is accompanied by a particle (WA), and the predicate has present tense,
   then   { indefinite (1, 0)   definite (1, 2)   generic (1, 3) }
   Example: INU-WA YAKUNITATSU(useful) DOUBUTSU(animal) DESU(is)
   _____Dogs[3] are useful animals._

4. When a noun is accompanied by a particle HE(to), MADE(up to) or KARA(from),
   then   { indefinite (1, 0)   definite (1, 2)   generic (1, 0) }
   Example: KARE-O(he) KUUKOU-MADE(airport) MUKAE-NI(to meet) YUKIMASHOO(let us go)
   _____Let us go to meet him at the airport._

There are many other expressions which give some clues for the referential property of nouns, such as (i) the noun itself, "CHIKYUU(the earth)" [definite], "UCYUU(the universe)" [definite], etc., (ii) nouns modified by a numeral (Example: KORE-WA(this) ISSATSUNO(one) HON-DESU(book)[indefinite]. (This is a book.)), (iii) the same noun presented previously (Example: KARE-WA(he) JOUYOUSHA(car)-TO(and) TORAKKU-O(truck) ICHIDAI-ZUTU(by ones) MOTTEIMASUGA(have), JOUYOUSHA-NIDAKE(car)[definite] HOKEN-O-KAKETEIMASU(be insured) (He has a car and a truck, but only the car is insured.)), (iv) adverb phrases, "ITSUMO(always)" , "NIHON-DEWA(in Japan)", etc. (Example: NIHON-DEWA SHASHOU-WA(conductor)[generic] JOUKYAKU(passenger)-NO(of) KIPPU-O(ticket) SIRABEMASU(check). (In Japan, the conductor checks the tickets of the passengers.)), (v) verbs, "SUKI(like)", 'TANOSHIMU(enjoy)", etc. (Example: WATASHI-WA(I) RINGO-GA(apple)[generic] SUKI-DESU(like). (I like apples.)).
   In the case of no clues, "definite" is given to a noun as a default value.

   Let us see an example which has several rule applications for the determination of the referential property of a noun. KUDAMONO(fruit) in the following sentence is an example.

 WAREWARE-GA(We) KINOU(yesterdav) TSUMITOTTA(picked) KUDAMONO-WA (fruit) AZI-GA(taste)
     IIDESU(be good).

                 The fruit that we picked yesterday tastes delicious.
   Seven rules are applied for the determination of the definiteness of this noun. These are the following.

 (i) When a noun is accompanied by WA, and the corresponding predicate has no past tense
     (KUDAMONO-WA AZI-GA IIDESU),
     then { indefinite (1, 0)   definite (1, 2)   generic (1, 3) }

---

[2] (a, b) means the _possibility_(a) _and_ the _value_(b).

[3] Both "a dog" and "the dog" are possible because of the generic subject.

(ii) When a noun is modified by an embedded sentence which has the past tense (TSUMITOTTA),
      then   { indefinite (1, 0)   definite (1, 1)   generic (1, 0) }

(iii) When a noun is modified by an embedded sentence which has a definite noun accompanied by WA
      or GA (WAREWARE-GA), then   { indefinite (1, 0)   definite (1, 1)   generic (1, 0) }

(iv) When a noun is modified by an embedded sentence which has a definite noun accompanied by a
      particle (WAREWARE-GA), then   { indefinite (1, 0)   definite (1, 1)   generic (1, 0) }

(v) When a noun is modified by a phrase which has a pronoun (WAREWARE-GA),
      then   { indefinite (1, 0)   definite (1, 1)   generic (1, 0) }

(vi) When a noun has an adjective as its predicate (KUDAMONO-WA AZI-GA <u>IIDESU</u>).
      then   { indefinite (1, 0)   definite (1, 3)   generic (1, 4) }

(vii) When a noun is a common noun (KUDAMONO),
      then   { indefinite (1, 1)   definite (1, 0)   generic (1, 0) }

As the result of the application of all these rules, we obtained the final score of { indefinite (1, 1)
definite (1, 9)   generic (1, 7) } for KUDAMONO, and "definite" is given as the decision.

4.2   Heuristic Rules for Number

1. When a noun is modified by SONO(its), ANO(that), KONO(this),
   then   { singular (1, 3)   plural (1, 0)   uncountable (1, 1) }
   Example: <u>ANO</u>(that) <u>HON-O</u> (book) KUDASAI (give me)
                       Give me <u>that book.</u>

2. When a noun is accompanied by a particle WA, GA, MO, O, and there is a numeral x which
   modifies the predicate of a sentence, and
   if   x = 1 ,   then   { singular (1. 2)   plural (1, 0)   uncountable (1, 0) }
   if   x $\geq$ 2 ,   then   { singular (1, 0)   plural (1, 2)   uncountable (1, 0) }
   Example: <u>RINGO-O</u>(apple) NIKO(two) TABERU(eat)
                       I eat two <u>apples</u>.

3. When a predicate, SUKI(like), TANOSHIMU(enjoy), etc. has a generic noun as an object, and the
   noun is accompanied by GA(for SUKI), or O(for TANOSHIMU),
   then { singular (1, 0)   plural (1, 2)   uncountable (1, 0) }
   Example: WATASHI-WA(I) <u>RINGO-GA</u>(apple) SUKI-DESU(like)
                       I like <u>apples</u>.

There are many other expressions which determine the number of a noun, such as (i) nouns modi-
fied by a numeral (Example: KORE-WA(this) ISSATSUNO(one) <u>HON-DESU</u>(book)[singular]. (This is
<u>a book</u>.)), (ii) verbs such as ATSUMERU(collect), AFURERU(be full with), (Example: WATASHI-WA(I)
NEKO-NO(about cat) <u>HON-O</u>(book)[plural] ATSUMETEIMASU(collect). (I collect <u>books</u> on cats.))
(iii) adverbs such as NANDO-DEMO(as many times as ...), IKURA-DEMO(as much ...) (Example:
RIYUU-WA(reason)[plural] IKURA-DEMO(as much …) SIMESEMASU(give). (I can give you a number
of reasons.)).

In the case of no clues, "singular" is given as a default value.

5   Experiments and Results

Experiments of the determination of the referential property and the number were done for the following three texts: typical example sentences in a grammar book "Usage of the English Articles" [2], the complete text of *a* Japanese popular folktale "The Old Man with a Wen" [5], a small fragment of an essay "TENSEI JINGO". The rules were written by referring to these sentences which have good established English translations. These sentences can be regarded as a training set. The results of the experiments are shown in Table 1. Here "correct" means that the result was correct. "Reasonable" means that the result is given, for example, as non-generic but the correct answer was definite, and so on. "Partially correct" means that the result was included in the correct answer. "Undecidable" means that we could not judge which category is correct by our linguistic intuitions. We obtained 85.5% success rate for the determination of the referential properties and 89.0% success rate for the numbers for all these learning samples. The scores of these tables show that the heuristic rules are well adjusted to these sentences, and are effective.

To testify the goodness of the rules we applied these heuristic rules to the following three other texts: a Japanese popular folktale "TURU NO ONGAESHI" [5], three small fragments of an essay "TENSEI JINGO" , "Pacific Asia in the Post-Cold-War World" (A Quarterly Publication of The International House of Japan Vol.12, No.2 Spring 1992). These test samples have good English translations. We used them to check the correctness of the results. The results are shown in Table 2. The success rates for the referential property and the number decreased down to 68.9% and 85.6% respectively by these test samples. These scores show, however, that the rules are *still* effective.

The success ratio will decrease greatly for the text areas which handle abstract notions such as philosophy and politics. We may have to change and increase heuristic rules for these text areas. At this moment we cannot say anything about whether we can write proper heuristic rules for such complex situations where delicate abstract notions are handled and the denotation is ambiguous.

As a conclusion we can say the following, There are of course many expressions and situations where inter-sentential information is necessary, but without utilizing it we can achieve a proper guess about the referential property and the number to a certain extent. By incorporating this mechanism into a machine translation system from Japanese into English we will be able to obtain better translation quality.

References

[1] Sadao Kurohashi: Syntactic Disambiguation using Evaluation Function for Case Structure, IPSJ-WGNL, 92-9, 1992 (in Japanese).

[2] Akihisa Kumayama: Usage of the English Articles, Taishukan Publisher, 1985 (in Japanese).

[3] Masayuki Ikeuchi: Meishiku-no Gentei Hyougen, New English Grammar Series, Vol. 6, Taishukan Publisher, 1985 (in Japanese).

[4] Kenkichirou Koizumi: Plurals and Articles in English, Japan Times, 1989 (in Japanese).

[5] Kiyoaki Nakao: The Old Man with a Wen, Eiyaku Nihon Mukashibanashi Series, Vol. 7, Nihon Eigo Kyouiku Kyoukai, 1985 (in Japanese).

Table 1: Learning sample

value	Referential property					Number				
	indef	def	gener	other	total	singl	plural	uncount	other	total
Usage of the English Articles(140 sentences, 380 nouns)										
correct	96	184	58	1	339	274	32	18	25	349
reasonable	0	3	1	0	4	1	1	1	0	3
partially correct	0	0	0	0	0	0	0	0	11	11
incorrect	4	25	7	1	37	3	10	0	4	17
% of correct	96.0	86.8	87.9	50.0	89 2	98.6	74.4	94 7	62.5	91.8
The Old Man with a Wen(104 sentences, 267 nouns)										
correct	73	140	6	1	222	205	24	5	0	234
reasonable	3	4	0	0	7	2	0	0	0	2
partially correct	0	0	0	0	0	0	0	0	7	7
incorrect	11	23	4	0	38	1	22	1	0	24
% of correct	83.9	84.0	60.0	100.0	83 2	98.7	52.2	83.3	0.0	87.6
an essay "TENSEI JINGO"(23 sentences, 98 nouns)										
correct	25	35	16	0	76	64	13	0	3	80
reasonable	0	4	2	0	6	2	1	0	0	3
partially correct	0	0	0	0	0	0	0	0	6	6
incorrect	5	10	1	0	16	1	6	1	1	9
% of correct	83.3	71.4	84.2	—	77 6	95.5	65.0	0.0	30.0	81.6
average										
% of appearance	29.1	57.7	12.8	0.4	100.0	74.2	14.6	3.5	7.7	100.0
% of correct	89.4	84.0	84.2	66.7	85 5	98.2	63.3	88.5	49.1	89.0

Table 2: Test sample

value	Referential property					Number				
	indef	def	gener	other	total	singl	plural	uncount	other	total
a folktale "TURU NO ONGAESHI"(263 sentences, 699 nouns)										
correct	109	363	13	10	495	610	13	1	1	625
reasonable	6	25	0	0	31	12	2	0	0	14
partially correct	0	0	0	0	0	0	0	0	1	1
incorrect	32	135	6	0	173	2	20	37	0	59
% of correct	74.2	69.4	68.4	100.0	70.8	97 8	37.1	2.6	50.0	89.4
an essay "TENSEI JINGO"(75 sentences, 283 nouns)										
correct	75	81	16	0	172	197	13	2	3	215
reasonable	8	9	1	0	18	3	1	0	0	4
partially correct	0	0	0	0	0	0	0	0	3	3
incorrect	33	51	9	0	93	3	55	3	0	61
% of correct	64.7	57.5	61.5	—	60.8	97 0	18.8	40.0	50.0	76.0
Pacific Asia in the Post-Cold-War World(22 sentences, 192 nouns)										
correct	21	108	11	2	142	157	6	1	1	165
reasonable	6	7	0	0	13	3	0	0	0	3
partially correct	0	0	0	0	0	0	0	0	0	0
incorrect	11	24	2	0	37	3	20	1	0	24
% of correct	55.3	77.7	84.6	100.0	74.0	96.3	23.1	50.0	100.0	85.9
average										
% of appearance	25.6	68.4	4.9	1.0	100.0	84.3	11.1	3.8	0.8	100.0
% of correct	68.1	68.7	69.0	100.0	68.9	97.4	24.6	8.9	55.6	85.6