Definiteness Predictions for Japanese Noun Phrases*

Julia E. Heine
Computerlinguistik
Universität des Saarlandes
66041 Saarbrücken
Germany
heine@coli.uni-sb.de

Abstract

One of the major problems when translating from Japanese into a European language such as German or English is to determine definiteness of noun phrases in order to choose the correct determiner in the target language. Even though in Japanese, noun phrase reference is said to depend in large parts on the discourse context, we show that in many cases there also exist linguistic markers for definiteness. We use these to build a rule hierarchy that predicts 79,5% of the articles with an accuracy of 98,9% from syntactic-semantic properties alone, yielding an efficient pre-processing tool for the computationally expensive context checking.

1 Introduction

One of the major problems when translating from Japanese into a European language such as German or English is the insertion of articles. Both German and English distinguish between the definite and indefinite article, the former, in general, indicating some degree of familiarity with the referent, the latter referring to something new. Thus by using a definite article, the speaker expects the hearer to be able to identify the object he is talking about, whilst with the use of an indefinite article, a new referent is introduced into the discourse context (Heim, 1982).

In contrast, the reference of Japanese noun phrases depends in large parts on the discourse context, taking a previous mention of an object and all properties that can be inferred from it, as well as world knowledge as indicators for definite reference. Any noun phrase whose referent cannot be recovered from the discourse context will in turn be taken as indefinite. However, noun phrases can also be explicitly marked for definiteness, forcing an interpretation of the referent independent of the discourse context. In this way, it is possible to trigger accommodation of previously unknown specific referents, or to get an indefinite reading even if an object of the same type has already been introduced.

For machine translation, it is important to find a systematic way of extracting the syntactic and semantic information responsible for marking the reference of noun phrases, in order to correctly choose the articles to be used in the target language.

For this paper, we propose a rule hierarchy for this purpose, that can be used as a pre-processing tool to context checking. All noun phrases marked for definiteness in any way are assigned their referential property, leaving the others underspecified.

After giving a short outline of related work in the next section, we will introduce our rule hierarchy in section 3. The resulting algorithm will be evaluated in section 4, and in section 5 we will address implementational issues. Finally, in section 6 we give a conclusion.

2 Related Work

The problem of article selection when translating from Japanese into any language requiring the use of articles has only been addressed systematically by a few authors.

(Murata and Nagao, 1993) define a heuristic rule base for definiteness assignment, consisting of 86 weighted rules. These rules use surface in-

* I would like to thank my colleagues Johan Bos, Björn Gambäck, Yoshiki Mori, Michael Paul, Manfred Pinkal, C.J. Rupp, Atsuko Shimada, Kristina Striegnitz and Karsten Worm for their valuable comments and support. This research was supported by the German Ministry of Education, Science, Research and Technology (BMBF) within the Verbmobil framework under grant no. 01 IV 701 R4.

formation in a sentence to estimate the referential property of each noun. During processing, each applicable rule assigns confidence weights to the three possible referential properties 'definite', 'indefinite' and 'generic'. These values are added up for each property, and the one with the highest score will be assigned to the noun in question. If no rule applies, the default value is 'indefinite'. This approach assigns the correct value in 85,5% of the cases when used with the training data, and 68,9% with unseen data.

(Bond et al., 1995) show how the percentage of noun phrases generated with correct use of articles and number in a Japanese to English machine translation system can be increased by applying heuristic rules to distinguish between 'generic', 'referential' and 'ascriptive' uses of noun phrases. These rules are ordered in a hierarchical manner, with later rules over-ruling earlier ones. In addition, for each noun phrase use there are specific rules, based on linguistic information, that assign definiteness to the noun phrases. Overall, in their system, insertion of the correct article can be improved by 12% yielding a correctness level of 77%.

In contrast to these approaches relying on monolingual indicators alone, (Siegel, 1996) proposes to assign definiteness during the transfer process. In a first stage, all lexically defined definiteness attributes are assigned. To all cases not covered by this, a set of preference rules is applied, if their translation equivalent in the target language is a noun. In addition to linguistic indicators from both the source and target language, the rules also take a stack of referents mentioned previously in the discourse into account. This combined approach is very successful, assigning the correct definiteness attributes to 98% of all relevant noun phrases in the training data.

In the approach described in the next section, we have taken up the idea of using both linguistic and contextual information for the assignment of definiteness attributes to Japanese noun phrases. However, instead of using merely a rule base, we propose a monotone algorithm based on a linguistic rule hierarchy followed by a context checking mechanism.

3 The Rule Hierarchy

The rule hierarchy we introduce in this paper has been devised from a systematic survey of some data from a Japanese corpus consisting of appointment scheduling dialogues.[1] Since dialogues in this domain tend to be short, on average consisting of just 14 utterances, most definite references have to be introduced by way of accommodation rather than referring back to the discourse context. Moreover, references to events have a particular tendency to be non-specific, i.e. stating their existence rather than explicating their identity. Non-specific references are by definition indefinite, whether the referent has been previously introduced to the context or not.

Neither accommodation nor non-specific reference can be realized without linguistic indicators, since they would otherwise interfere with the context-based distinction between definite and indefinite reference within a discourse. The appointment scheduling domain is therefore ideal for a case study aimed at extracting linguistic indicators for definiteness.

3.1 Overview

Explicit marking for definiteness takes place on several syntactic levels, namely on the noun itself, within the noun phrase, through counting expressions, or on the sentence level. For each of these syntactic levels, a set of rules can be defined by generalizing over the linguistic indicators that are responsible for the definiteness attributes carried by the noun phrases in the corpus. Each of these rules consists of one or more preconditions, and a consequent that assigns the associated definiteness attribute to the respective noun phrase when the preconditions are met.

As it turns out, none of the rules defined on the same syntactic level interfere with each other, since they either assign the same value, or their preconditions cannot possibly be met at the same time. Thus the rules can be grouped together into classes corresponding to the four

[1] In this survey, all the noun phrases from 10 dialogues were analyzed in detail, determining the regularities that led to definiteness predictions. These were then formulated into a set of rules and arranged in a hierarchical manner to rule out wrong predictions. A more detailed description of the methods used and a full list of the rules can be found in (Heine, 1997).

syntactic levels they are defined on. There is a clear hierachy between the four classes, with all rules of one class given priority over all rules on a lower level, as shown in figure 1. Note that even though the rule classes are defined in terms of syntactic levels, the sequence of rule classes in our hierarchy does not correspond in any way to syntactic structure.

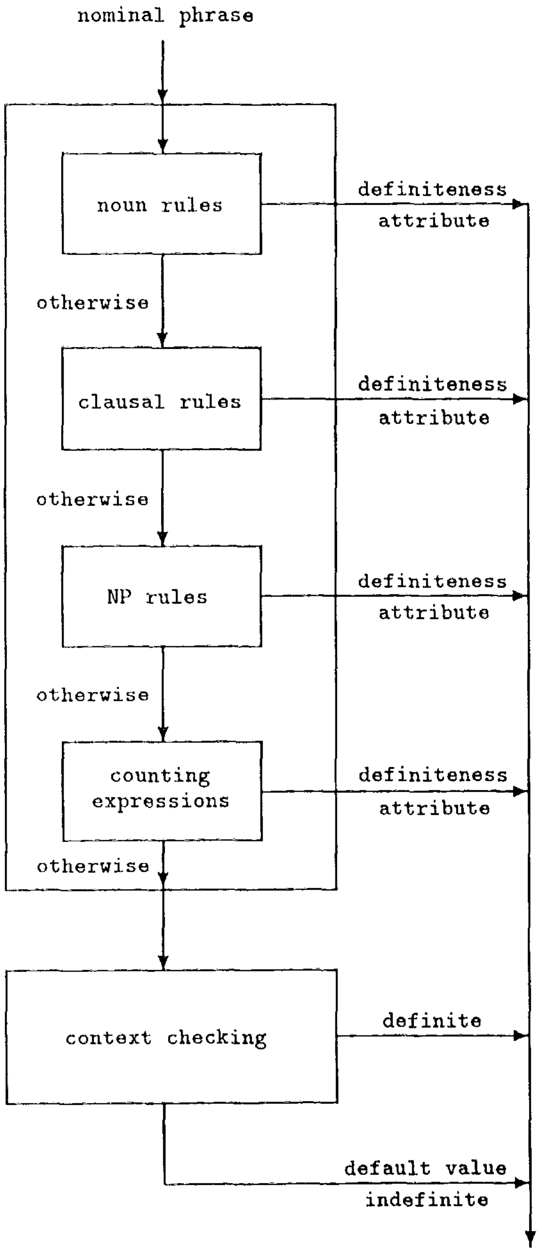

Figure 1: Definiteness Algorithm

3.2 Noun rules

On the noun level, the lexical properties of the noun or one of its direct modifiers can determine the reference of the noun in question.

There are a number of nouns, that can be marked as definite on their lexical properties alone, either because they refer to a unique referent in the universe of discourse, or because they carry some sort of indexical implications. The referent is thus described uniquely with respect to some implicitly mentioned context. For example, there exist a number of nouns that implicitly relate the referent with either the hearer or the speaker, depending on the presence or absence of honorifics[2], respectively. In the appointment scheduling domain, the most frequently used words of this class are *(go)yotei* (your/my schedule), *(o)kangae* (your/my opinion) and *(go)tsugoo* (for you/me).

Indexical time expressions like *konshuu* (this week) or *raigatsu* (next month) refer to a specific period of time that stands in a certain relation to the time of utterance. Even though they do not necessarily have to stand with an article in the target language, the reference is still definite, as in the following example:

(1) **raishuu** *desu* *ne*
next week to be isn't it

'That is **(the) next week**, isn't it?'

The interpretation of a modified noun is typically restricted to a specific referent by the modification, thus making it definite in reference. Restrictive modifiers of this type are, for example, specifiers like demonstratives and possessives, as well as time expressions and attributive relative clauses, as shown in the following examples.

(2) **tooka no** *shuu desu*
tenth GEN week to be

'That is **the week of the tenth**.'

(3) **nijuurokunichi kara hajimaru**
twentysixth from to begin
shuu wa ikaga deshoo ka
week TOPIC how to be QUESTION

[2] In Japanese, there are two honorific prefixes, *go* and *o*, that can be used to politely refer to things related to the hearer. However, there are no such prefixes to humbly refer to things relating to oneself.

'How is **the week** beginning the 26th?'

However, indefinite pronouns, as for example *hoka* (another), also fall into the category of modifiers, but explicitly assign indefinite reference to the noun they modify. These are usually used to introduce a new referent into a context already containing one or more referents of the same type.

(4) **hoka no** *hi* *erabashite* *itadaite* *mo*
different day choose receive also
ii *n desu ga*
good DISCREL

'Could I ask you to choose a different day?'

At present, there are nine rules belonging to the noun class, only one of which assigns indefinite reference whilst all others assign definite reference to the noun in question.

3.3 Clausal rules

On the sentence level, verbs may carry strong preferences for the definiteness of one or more of their arguments, somewhat in the way of domain specific patterns. Generally, these patterns serve to specify whether a complement to a certain verb is more likely to be definite or indefinite in a semantically unmarked interpretation. For example, in a sentence like 5, *kaigi ga haitte orimasu* corresponds to the pattern 'EVENT *ga hairu*' ('have an EVENT scheduled'), where the scheduled event denoted by EVENT is indefinite for the unmarked reading.

(5) *kayoobi* *wa* *gogo* *sanji* *made*
Tuesday TOPIC pm 3 o'clock until
kaigi **ga** **haitte** *orimasu* *node*
meeting NOM have scheduled since

'since I have **a meeting** scheduled until 3 pm on Tuesday'

On the other hand, in sentence 6, *kaigi ga owarimasu* is an instance of the pattern 'EVENT *ga owaru*' ('the EVENT will end'), where, in the unmarked reading, the event that ends is presupposed to be a specific entity, whether it is previously known or not.

(6) *juuniji* *ni* **kaigi** **ga**
12 o'clock at meeting NOM
owarimasu *node*
to end since

'since **the meeting** will end at 12 o'clock'

The object of an existential question or a negation is by default indefinite, since these sentence types usually indicate the (non)existence of the noun in question. Thus, for example, in the two sentence patterns 'X *wa arimasu ka*' ('Is there an X?') and 'X *wa arimasen*' ('There is no X.') the object instantiating X is indefinite, unless marked otherwise.

In addition to these sentence patterns, there are a number of nouns that can be followed by the copula *suru* to form a light verb construction. These constructions usually come without a particle and are treated as compound verbs, as for example *uchiawase suru* ('to arrange'). However, these nouns can also occur with the particle *o*, as in *uchiawase o suru*, introducing an ambiguity whether this expression should be treated as a light verb construction or as a normal verb complement structure. Since this ambiguity can best be resolved at some later point, the noun should be marked as being indefinite, irrespective of whether it will eventually be generated as a noun or a verb in the target language.

(7) *raishuu* *ikoo de*
next week from ... onwards
uchiawase **o** *shitai*
arrangement ACC want to make
n desu ga
DISCREL

'I would like to make **an arrangement** from next week onwards'

To override any of these default values, the noun will have to be explicitly marked, using any of the markers on the noun level. Thus we take the clausal rules to be between the top level noun rules and all other rules further down the hierarchy.

From the appointment scheduling domain, eight sentence patterns were extracted, where six assign the default indefinite and two indicate definite reference. Thus, together with the light verb constructions, there are nine rules in this class.

3.4 Noun phrase rules

The postpositional particles that complete a noun phrase in Japanese serve primarily as case markers, but can also influence the interpretation of the noun with respect to definiteness. However, the definiteness predictions triggered by the use of particles can be fairly weak and are easily overridden by other factors, thus placing the rules emerging from these patterns near the bottom of the hierarchy.

The main postpositions indicating definite reference are the topicalization particle *wa* in its non-contrastive use[3], the boundary markers *kara* (from) and *made* (to) and the genitive marker *no*, especially in conjunction with *hoo* (side), as indicated by the following examples.

(8) *chotto* **idoo** **no** *jikan*
unfortunately transfer GEN time
ga *torenaiyoo* *desu ne*
NOM take not DISCREL

'Unfortunately, there is no time for **the transfer.**'

(9) *genkoo* **no** **hoo** *mada* *tochuu*
manuscript GEN side not yet ready
dankai *desu* *keredomo*
state to be DISCREL

'**The manuscript** is not ready yet.'

All of the four noun phrase rules in the current framework indicate definite reference.

3.5 Counting expressions

As it turns out, there is one more level to the rule hierarchy. Even though counting expressions are semantically modifiers, they do not syntactically modify the noun itself but rather the entire noun phrase. They do not have to be adjacent to the noun phrase they modify, since they are marked by a counting suffix indicating the type of objects counted.

(10) *nijuuhachinichi* *ga* *gogo* *ni*
twentyeighth NOM afternoon in
kaigi *ga* **ikken** *haitte orimasu*
meeting ACC one be scheduled

'There is **one/a meeting** scheduled on the twentyeighth.'

Semantically, counting expressions imply the existence of a certain number of the objects counted, in the same way that the indefinite article does. These expressions are therefore taken to be indefinite by default, but can be made definite by any of the other rules. Counting expressions thus make up a class of their own on the lowest level of the hierarchy.

3.6 Underspecified values

As might be expected from the concept of preprocessing, there will be a number of noun phrases that cannot be assigned a definiteness attribute by any of the rules described above. These will remain underspecified for definiteness until an antecedent can be found for them by the context checking mechanism, or until they are assigned a default value.

By introducing a value for underspecification, it is possible to postpone the decision whether a noun phrase should be marked definite or indefinite, without losing the information that it must be marked eventually. Since default values are only introduced when a value is still underspecified after the assignment mechanism has finished, there is no need to ever change a value once it has been assigned. This means, that the algorithm can work in a strictly monotone manner, terminating as soon as a value has been found.

4 Evaluation

4.1 Performance of the algorithm

The performance of our framework is best described in terms of recall and precision, where recall refers to the proportion of all relevant noun phrases that have been assigned a correct definiteness attribute, whilst precision expresses the percentage of correct assignments among all attributes assigned.

The hierarchy was designed as a pre-process to context checking, extracting all values that can be assigned on linguistic grounds alone, but leaving all others underspecified. It is therefore

[3] This means, that definite reference is indicated by the main use of the particle *wa*, namely as a topic marker, stressing the discourse referent the conversation is about. There is another, contrastive use of *wa*, which introduces something in contrast to another discourse referent. Naturally, this use may introduce a related, albeit previously unknown — and thus indefinite — referent.

	noun rules	clausal rules	NP rules	count rules	total
occurrences	159	62	53	1	275
correct	158	60	53	1	272
incorrect	1	2	0	0	3
precision	99,4%	96,8%	100%	100%	**98,9%**

Table 1: Precision of the rules

to be expected that its coverage, i.e. the percentage of noun phrases assigned a value by the hierarchy, is relatively low. However, since we propose that the decision algorithm should be monotone, it is vitally important for the precision to be as near to 100% as possible. Any wrong assignments at any stage of the process will inevitably lead to incorrect translation results.

To evaluate the hierarchy, we tested the performance of our rule base on 20 unseen dialogues from the corpus. All noun phrases in the dialogues were first annotated with their definiteness attributes, followed by the list of rules with matching preconditions. As a second step, the rules applicable to each noun phrase were ordered according to their class, and the prediction of the one highest in the hierarchy was compared with the annotated value.

In the test data, there are 346 noun phrases that need assignment of definiteness attributes.[4] Table 1 shows the number of noun phrase occurrences covered by each rule class, i.e. the number of times one of the noun phrases was assigned a definiteness attribute by any of the rules from each class. This value was then further divided into the number of correct and incorrect assignments made. From this, the precision was calculated, dividing the number of values correctly assigned by the number of values assigned at all. Overall, with a precision of 98,9%, the aim of high accuracy has been achieved.

Dividing the number of correct assignments by the number of noun phrases that need assignment, we get a recall of 78,6%. Thus, within the appointment scheduling domain, the hierarchy already accounts for 79,5% of all relevant noun phrases, leaving just 20,5% for the computationally expensive context checking.

Of the 71 noun phrases left underspecified, 40 have definite reference, suggesting 'definite' as the default value if the hierarchy was to be used as the sole means of assigning definiteness attributes. This means, that a system integrating this algorithm with an efficient context checking mechanism should have a recall of at least 90%, since this is what can already be achieved by using a default value.

4.2 Comparison to previous approaches

The performance of our framework has been found to be better than both of the heuristic rule based approaches introduced in section 2, even before context checking. However, our framework was defined and tested on the restrictive domain of appointment scheduling. Most of the really difficult cases for article selection, as for example generics, do not occur in this domain, whilst both (Murata and Nagao, 1993) and (Bond et al., 1995) build their theories around the problem of identifying these. There are no statistics on the performance of their systems on a corpus that does not contain any generics.

The transfer-based approach of (Siegel, 1996) also covers data from the appointment scheduling domain, using both linguistic and contextual information for assigning defininteness. However, her results can still not be compared with our approach, since we do not have any figures on how high the recall of our algorithm is with context checking in place. In addition, the performance data given for our hierarchy was derived from unseen data rather than the data that were used to draw up the rules, as in Siegel's case.

[4] Additionally, there are 388 time expressions (i.e. dates, times, weekdays and times of day) that under certain conditions also need an article during generation. However, these were excluded from the statistics, since nearly all of them were found to be trivially definite, somehow artificially pushing the recall of the rules in the hierarchy up to 88,8%.

Even though no direct comparison is possible because of the different test methods and data sets used, we have been able to show that an approach using a monotone rule hierarchy that can be easily integrated with a context checking mechansim leads to very good results.

5 Implementation

The current framework has been designed as part of the dialogue and discourse processing component of the Verbmobil machine translation system, a large scale research project in the area of spontaneous speech dialogue translation between German, English and Japanese (Wahlster, 1997). Within the modular system architecture, the dialogue and discourse processing is situated in between the components for semantic construction (Gambäck et al., 1996) and semantic-based transfer (Dorna and Emele, 1996). It uses context knowledge to resolve semantic representations possibly underspecified with respect to syntactic or semantic ambiguities.

At this stage, all the information needed for definiteness assignment is easily accessible, enabling the rules in our hierarchy to be implemented one-to-one as simple implications. Since all information is accessible at all times, the application of the rules can be ordered according to the hierarchy. Only if none of the rules given in the hierarchy are applicable, will the context checking process be started. If an antecedent can be found for the relevant noun phrase, it will be assigned definite reference, otherwise it is taken to be indefinite.

The algorithm will terminate as soon as a value has been assigned, thus ensuring monotonicity and efficiency, as 45% of all noun phrases are already assigned a value by one of the noun rules at the top of the hierarchy.

6 Conclusion

In this paper, we have developed an efficient algorithm for the assignment of definiteness attributes to Japanese noun phrases that makes use of syntactic and semantic information.

Within the domain of appointment scheduling, the integration of our rule hierarchy reduces the need for computationally expensive context checking to 20,5% of all relevant noun phrases, as 79,5% are already assigned a value with a precision of 98,9%.

Even though the current framework is to a large extent domain specific, we believe that it may be easily extended to other domains by adding appropriate rules.

References

Francis Bond, Kentaro Ogura, and Tsukasa Kawaoka. 1995. Noun phrase reference in Japanese-to-English machine translation. In *Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 1–14.

Michael Dorna and Martin C. Emele. 1996. Semantic-based transfer. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 1, pages 316–321, København, Denmark. ACL.

Björn Gambäck, Christian Lieske, and Yoshiki Mori. 1996. Underspecified Japanese semantics in a machine translation system. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation*, pages 53–62, Seoul, Korea.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases.* Ph.D. thesis, University of Massachusetts.

Julia E. Heine. 1997. Ein Algorithmus zur Bestimmung der Definitheitswerte japanischer Nominalphrasen. Diplomarbeit, Universität des Saarlandes, Saarbrücken. available at: http://www.coli.uni-sb.de/~heine/arbeit.ps.gz (in German).

Masaki Murata and Makoto Nagao. 1993. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 218–225.

Melanie Siegel. 1996. Preferences and defaults for definiteness and number in Japanese to German machine translation. In Byung-Soo Park and Jong-Bok Kim, editors, *Selected Papers from the 11th Pacific Asia Conference on Language, Information and Computation.*

Wolfgang Wahlster. 1997. Verbmobil - Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. Verbmobil Report 198, DFKI GmbH. (in German).