Preferences and Defaults for Definiteness and Number in Japanese to German Machine Translation

Melanie Siegel
DFKI GmbH
Stuhlsatzenhausweg 3, 66123 Saarbrücken
siegel@dfki.uni-sb.de
tel.0681-302-5284
fax 0681-302-5338

Abstract

A significant problem when translating Japanese dialogues into German is the missing information on number and definiteness in the Japanese analysis output. The integration of the search for such information into the transfer process provides an efficient solution. General transfer rules, preference rules and default rules are combined. The transfer includes conditions to make it possible to consider external knowledge. Thereby, grammatical and lexical knowledge of the source language, knowledge of lexical restrictions on the target language, domain knowledge and discourse knowledge are accessible.

1 Introduction

One of the significant problems in Japanese to German machine translation is that information on definiteness and number is in most cases not available on the surface of the Japanese utterance. Japanese has neither number agreement between verbs and nouns nor obligatory plural morphemes. Optional plural morphemes like *tachi* or *domo* are only available for nouns that refer to persons. However, for the generation of German utterances the generator needs such information, as in many cases determiners are obligatory. We analyzed the problem based on empirical material collected in the domain of appointment scheduling. In our setting, an interpreter translated 10 dialogues of Japanese and German speakers. Consider the following example from our data:

Japanese:

kayoobi	wa	watashidomo	no	tokoro	de
Tuesday	TOPIC	we	GEN	side	CASE DE

wa	*kyuujitsu*	na	no	de	tabun	*kaigi*
TOPIC	holiday		copula		maybe	meeting

ni	sanka	suru	koto	wa	dekimasen
CASE NI	participate	do	NOM	TOPIC	cannot

German translation:

auf	unserer	Seite	ist	Freitag	*ein*	*Feiertag*	vielleicht
on	our	side	is	Friday	a	holiday	maybe

können	wir	an	*dem*	*Treffen*	nicht	teilnehmen
can	we	at	the	meeting	not	participate

(On our side Friday is a holiday and we maybe cannot participate in the meeting)

The information that *Feiertag* has to be preceded by the singular indefinite (masculine) determiner *ein* and that *Treffen* has to be preceded by the singular definite (neuter) determiner *dem* does not come out of the surface of the Japanese utterance and therefore cannot be included in the parsing result. It is not an adequate solution to transfer an underspecified representation to the German generation module, because the information that is needed to decide on the definiteness and number of the noun phrase partly comes out of the Japanese surface, partly out of German lexical restrictions and partly out of domain and discourse restrictions. Not all of this information is available in the generation phase. We argue that it is an interlingual problem and therefore must be solved in the transfer module.

2 Previous approaches

[Murata and Nagao, 1993] describe a solution model that searches for information at the surface of the Japanese utterances to give hints for the choice of determiners in the English counterpart. They state heuristics for the possibility of definiteness and number values concerning words or constructions in the Japanese utterances. Such information can be determiners in Japanese or particle and tense information.

But this is only one of the relevant aspects, because it does neither consider restrictions on definiteness and number coming from the target language, nor restrictions based on domain and discourse. Just as little as it is an inherent German problem it is an inherent Japanese problem.

[Bond *et al.*, 1994] already state that the inclusion of information about the target language (English in their case) increases the rate of correct translations. They enrich the heuristics with information on countability of English nouns. But their approach still lacks the integration of knowledge about discourse and domain, which is relevant as we will show.

Every approach that is external to the transfer process cannot include source and target language information at the same time and be thus effective. It is essential to find a mechanism that makes it possible to consider discourse and

domain knowledge.

3 Transfer rules

Only in some cases, where definite articles, quantifying adjectives or certain genitive constructions can be found in the Japanese source language utterance, the parsing result can directly contain information on number and definiteness. The information that is necessary to generate the German determiners has to be searched in the surface of the Japanese utterance (as it is described by [Murata and Nagao, 1993]), in lexical information of the German target language (as it is described by [Bond *et al.*, 1994]) and in external knowledge sources.

The solution of our approach is based on the idea of transfer-based machine translation. The representation for the transfer rules are expressions in simplified RQLF-format [Alshawi, 1992]. An RQLF, Resolved Quasi-Logical Form, is an underspecified semantic representation that includes context information. Nouns with determiners are represented as *qterms*. An example is the representation for *a house*:

$$qterm = (< t = quant, p = indef, n = sing, l = a >, house)$$

t is the type of expression, p the phrase type, n contains number information and l the lexical realization.

The integration of the search for definiteness and number in the transfer process reduces the complexity of the problem, because it is possible to state a number of transfer rules without searching for information on number and definiteness. Another advantage is that no extra process has to be activated to find information on the Japanese sentence surface. A practical aspect is therefore the avoidance of redundancy in the translation process, contrasting Murata/Nagaos approach. Only when no transfer rule can be found that directly gives information on definiteness and number, preference rules are activated to search for the missing information. The rule format of the transfer rules is the following:

$$transfer(JapaneseRQLF \Longrightarrow GermanRQLF) : -conditions.$$

It contains the Prolog predicate $transfer$, a translation rule expressing the relation between the source and target language expression and (optionally) one or more conditions. The RQLF expressions can be complex, including for example the representation of a determiner and the corresponding noun. They can include variables. Conditions are optional Prolog clauses. They can restrict the transfer rule to a certain value in the RQLF or to a speechact. This makes it possible to include domain knowledge. They can also be $transfer$-predicates so that the rule is recursive. This is needed for the case that a rule inserts information on only number *or* definiteness and another one is needed to insert the missing information. Another possibility is the condition $definiteness$ that

looks for further information on definiteness after the information on number was found. These conditions are responsible for discourse knowledge. A combination of conditions is also allowed. The searching strategy of the transfer rules is determined by the Prolog mechanism of backtracking.

3.1 Rules that avoid the necessity to insert information on number and definiteness

In many cases a preferred German equivalent does not contain a noun, and therefore no more information on number and definiteness is needed. These are — on the one hand — general translation equivalents for complex expressions and — on the other hand — realizations of speechacts in the domain. Examples for the first ones are:

- *hayai jikan – früh* (early time – early)
- *nagai jikan – lange* (long time – long)
- *yasumi – geschlossen* (holiday – closed)

Such general translation equivalents are easy to state, as for example *nagai jikan – lange*:

$$transfer([jikan1,nagai1] \Longrightarrow [lang\_\ prop]).$$

Temporal expressions are translated stereotypically without searching for information on number and definiteness, as for example *niji ni – um zwei Uhr* (at one o'clock). Some Japanese noun phrases that contain two nouns connected by a genitive *no* can have German equivalents that contain only one noun: *watashidomo no tokoro – wir* (our side – we). Other Japanese noun phrases containing no-phrases have a German equivalent with a nominal compound: *getsuyoobi no gogo – Montag Nachmittag* (afternoon of Monday – Monday afternoon). The German nominal compound gets only one determiner; as soon as a restriction for one of the parts is found, the determiner can be decided on. Other cases are domain-specific: Japanese *mina* in our data is always translated as *alle Mitarbeiter* (all staff-members), but could be – for example in another domain – *alle Studenten* (all students). It always (independent of the domain) has to be translated with plural. The rule in our domain is:

$$transfer([mina] \Longrightarrow$$
$$[qterm = [t = quant, p = def, n = plural, l = alle], Mitarbeiter]).$$

An example, where the number information is dependent on the domain is the following:

Japanese:

kono	hi	wa	shain	wa	kimasen
this	day	TOP	staff-member	TOP	do not come

German:

An	diesem	Tag	kommen	die	Mitarbeiter
on	this	day	come,pl.	the,pl.	staff-members
nicht					
not					

or:

An	diesem	Tag	kommt	der	Mitarbeiter
on	this	day	come,sg	the,sg.	staff-members
nicht					
not					

The translation of *shain* depends on if the company has one or more staff-members. It could be thought of adding a condition $domain(D)$, if the rules are used in a system that concerns more than one domain.

Strictly speaking, information on gender has to be found, too. But this is an inherent German problem and underlies German lexical restrictions and domain restrictions and is therefore left to generation. Examples for speechact realizations are *yotei wo tatetai – schlage ich vor* (I would like to set up the plan – I propose) and *donna yotei ni naru ka – wie ist...?* (what plan will become – how about...?). By using indicators for speechacts one can try to find pragmatic translation equivalents instead of literal ones: Both examples belong to the speechact *proposal*. Thus, transfer rules can include a condition that is an inquiry to the dialogue component. The project VERBMOBIL defines dialogue acts for the appointment scheduling domain, which could also be used for this purpose. See [Jekat *et al.*, 1995] for further information.

The empirical base is built by 10 collected dialogues of appointment scheduling between German and Japanese speakers that were translated by an interpreter. The data includes 566 noun tokens. 18.9% of these Japanese nouns have German equivalents that are not nouns. 34.63% are temporal expressions that are translated stereotypically. That makes more than 50% of all nouns where neither search for information on number nor for such on definiteness is necessary when first adequate transfer rules — general ones or domain specific ones — are searched for. This already is a strong argument to integrate the solution of the problem into the transfer process and to adapt the transfer to the domain.

3.2 General rules

Numerals in Japanese give clear information on number and have an unambiguous German translation, as for example:

- *ichijikan, sanjikan – eine Stunde, drei Stunden* (one hour, three hours)
- *hitori, futari – eine Person, zwei Personen* (one person, two persons)
- *hitori no hito, yonin membaa – eine Person, vier Mitglieder* (one person, four members)
- *kenkyuuin no hitori – einer unserer Forscher* (one of our researchers)

These cases underly general transfer rules for number. Still, information on definiteness has to be found, as the following transfer rule shows:

$$transfer([sanjikan] \Longrightarrow$$
$$[qterm = [t = quant, p = P, n = plural, l = drei], Stunden]) :-$$
$$definiteness(sanjikan, P).$$

It translates *sanjikan* into *drei Stunden* or *die drei Stunden*. The condition $definiteness$ is a predicate to test whether an entity is pre-mentioned (that is, included on a stack of pre-mentioned entities) and thus definite, or not pre-mentioned and thus indefinite. In our implementation, the stack is not only used for determining definiteness, but also for resolving zero pronominals (see [Siegel, 1996] for further information). These cases concern 7.42% of the nouns in our data base.

In some cases Japanese nominal phrases contain determiners, as *kono jikantai* (this period of time/these periods of time) and *sono jikan* (that time/that times). In these cases the translation concerning definiteness is straightforward:

$$transfer($$
$$[qterm = [t = quant, p = def, n = N, l = kono], X] \Longrightarrow$$
$$[qterm = [t = quant, p = def, n = ND, l = dies], XD]) :-$$
$$transfer\ (X, XD).$$

A definite ($p = def$) Japanese nominal phrase with a determiner *kono* and without information on number ($n = N$) is transferred to a definite German nominal phrase with a determiner *dies* (this) and German number information ND. The call $transfer(X, XD)$ initiates the search for a transfer equivalent of the noun phrase and its number information and unifies the result with the found $qterm$. But not only determiners lead to a situation where transfer rules concerning definiteness can be stated straightforwardly. Other possibilities are some kinds of adjectives and genitive constructions, as for example:

- *onaji shuu – dieselbe Woche/dieselben Wochen* (the same week(s))
- *tsugi no hi – der nächste Tag/die nächsten Tage* (the next day(s))
- *kondo no kaigi – das nächste Treffen/die nächsten Treffen* (the next meeting(s))

All of these require a translation with definite determiner.

3.3 Preference rules and the default

[Schmitz and Quantz, 1993] present an hybrid model for the search for information that combines exact knowledge with default knowledge. Default knowledge can be formulated as valid in a domain. Some entities in a domain are known and unique. These have to be translated singular and definite. Those are in our domain, for example, days of the week, the lunch break and the meeting. It is necessary to include domain-specific default transfer rules, as for example:

$$transfer(kaisha \Longrightarrow \\ [qterm = [t = quant, p = def, n = sg, l = L], Firma]).$$

Pre-mentioned entities have to be kept in a stack and have to be translated with definite determiner. An option with strong preference is to copy the information on number from the previous mentioned entity, as in the following example:

Japanese:

kono	jikantai	wa	dekireba	sakete
this	period(s) of time	TOP	if possible	keep free
itadakitai	to	omoimasu		
HON	COMPL	think		

German:

Ich	denke,	daß	ich	diesen Zeitraum	/
I	think	that	I	this period of time	/
diese Zeiträume		möglichst	freihalten	möchte	
these periods of time		if possible	keep free	want	

The translation depends on if it was spoken about one or more periods of time before.

In German sentences with copula predicates number agreement between subject and x-complement is preferred. This can be stated as a preference rule. This is not a general rule, as the example *wir sind ein Projektteam* (we are a project team) shows. Consider the following examples:

Japanese:

getsuyoobi	wa	kyuujitsu	desu
Monday	TOP	holiday	is

German:

Montag	ist	ein	Feiertag
Monday	is	a	holiday

and

Japanese:

getsuyoobi	to	kayoobi	wa	kyuujitsu	desu
Monday	and	Tuesday	TOP	holiday	is

German:

Montag	und	Dienstag	sind	Feiertage
Monday	and	Tuesday	are	holidays

The transfer rule is as follows:

$transfer($
$[desu, QTERMJ, ARG1J, ARG2J] \Longrightarrow$
$[sein, QTERMD, ARG1D,$
$[qterm = [t = quant, p = indef, n = N, l = L]], ARG2D]) : -$

$transfer(ARG1J, ARG1D),$ $\%transfer\ the\ first\ argument$
$value(qterm/n, ARG1D, N),$ $\%copy\ information\ on\ number$
$transfer(QTERMJ, QTERMD).$

If no transfer rule or preference rule is applicable, singular indefinite is inserted as a default. The analysis of the data shows that in most cases that do not fall under the categories described above, this default leads to a correct translation.

The preferences for number are:

1. Number information that comes from the Japanese nouns is copied into the German *qterm*.
2. Known and unique entities in the domain are translated with singular determiner.
3. Number agreement in copula sentences.
4. Default: Singular.

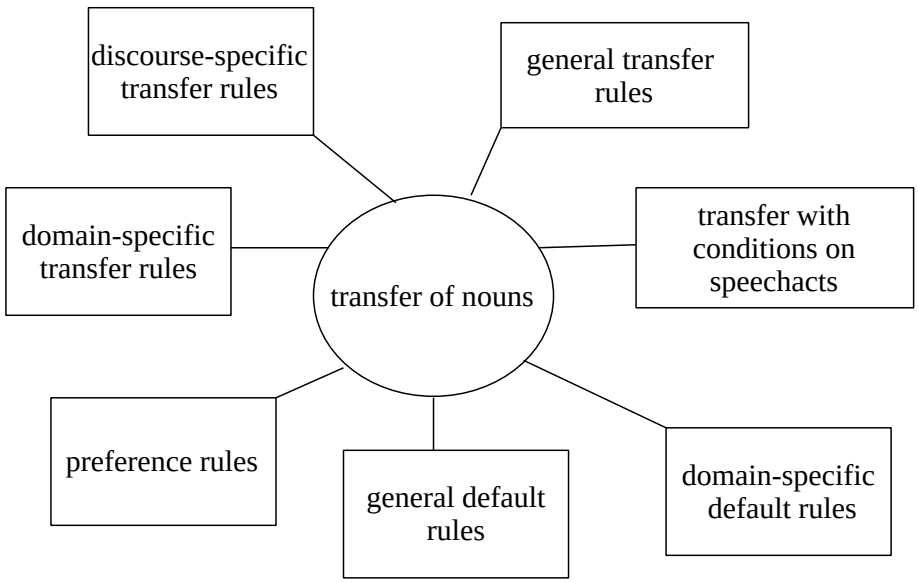
The preferences for definiteness are:

1. Known and unique entities in the domain are translated with definite determiner.
2. Noun phrases with *kono, sono, dono, onaji no, tsugi no,* or *kondo no* in a nominal phrase are translated definite.
3. Pre-mentioned entities are translated with definite article.
4. Default: Indefinite.

4 Summary

The problem of missing number and definiteness in translating Japanese nouns into German is significant as it occurs with every Japanese utterance that has to be translated. To solute the problem it is necessary to combine knowledge on the Japanese source language with such on the German target language and discourse and domain knowledge. Previous approaches lack this possibility to connect the different knowledge sources.

We integrate preference rules and default rules concerning number and definiteness into the transfer and are therefore able to consider source and target language at the same time. Domain and discourse knowledge can be referred to by conditions that restrict the transfer rules. Preference rules are stated for domain-specific and discourse knowledge. Domain-specific knowledge is encoded in a stack of unique entities of a domain. Discourse knowledge is also encoded in a stack on pre-mentioned entities. The following picture shows the connection of different kinds of transfer rules to translate Japanese to German noun phrases.



We have shown that combinating transfer and the search for information on number and definiteness reduces the problem to a reasonable extend. It can be shown that though general rules have to be preferred to domain-specific ones, domain-specific rules play an important role in translating Japanese noun phrases into German.

The described transfer rules and preference restrictions are based on observations on a corpus. They are implemented in a Prolog program.

References

[Alshawi, 1992] H. Alshawi. *The Core Language Engine.* Cambridge: The MIT

Press, 1992.

[Bond *et al.*, 1994] F. Bond, K. Ogura, and S. Ikehara. Countability and number in Japanese to English machine translation. In *Proceedings of Coling '94*, pages 32–38, 1994.

[Jekat *et al.*, 1995] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J.Joachim Quantz. Dialogue acts in VERBMOBIL. Verbmobil-Report 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.

[Murata and Nagao, 1993] M. Murata and M. Nagao. Determination of referential property and number of nouns in Japanese sentences for machine translation into English. In *Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 218–225, 1993.

[Schmitz and Quantz, 1993] B. Schmitz and J. Quantz. Defaults in machine translation. KIT-Report 106, Technische Universität Berlin, 1993.

[Siegel, 1996] Melanie Siegel. *Die maschinelle Übersetzung aufgabenorientierter japanisch-deutscher Dialoge. Lösungen für Translation Mismatches.* Ph.D.Thesis, Universität Bielefeld, 1996.