Definiteness and Number in Japanese to German Machine Translation

Melanie Siegel

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. Transfer rules, preference rules and default rules are combined. Thereby, grammatical and lexical knowledge of the source language, knowledge of lexical restrictions on the target language, domain knowledge and discourse knowledge are accessible.

Eins der signifikanten Probleme in der maschinellen Übersetzung japanischer in deutsche Sprache ist die fehlende Information über Numerus und Definitheit im japanischen Analyse-Output. Eine effiziente Lösung dieses Problems ist es, die Suche nach der relevanten Information in den Transfer zu integrieren. Transferregeln werden mit Präferenzregeln und Default-Regeln kombiniert. Dadurch wird Information über lexikalische Restriktionen der Zielsprache, über die Domäne und über den Diskurs zugänglich.

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. However, for the generation of German utterances the generator needs such information as in many cases determiners are obligatory. Consider the following example from our collected 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.

[MN93] describe a solution model that uses heuristical methods to search for information only at the surface of the Japanese utterance to give hints for the choice of determiners in the English counterpart. But this is only one of the relevant aspects, because just as little as it is an inherent German problem it is an inherent Japanese problem. [BOI94] already state that the inclusion of information about the target language (English in their case) increases the rate of correct translations. But their approach still lacks integration of knowledge on discourse and domain, which is relevant as we will show. Our approach goes further: We integrate the resolution into the transfer process. Furthermore we show that integration of discourse and domain knowledge is essential. This knowledge is encoded in preference rules and default rules.

2 Transfer Rules

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 [Als92][1]. 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 the search 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 give 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. 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. A combination of conditions is also allowed. The searching strategy of the transfer rules is determined by the Prolog mechanisms.

2.1 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) and *yasumi* – *geschlossen* (holiday – closed). Such general translation equivalents are easy to state, as for example *nagai jikan* – *lange*:

[1] An RQLF, Resolved Quasi-Logical Form, is an underspecified semantic representation that includes context information

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

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 researchers), but could be – for example in another domain – *alle Studenten* (all students). 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 a predicate to determine speechacts.

18.9% of the Japanese nouns in our data[2] 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.

2.2 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) and *kenkyuuin no hitori – einer unserer Forscher* (one of our researchers). These cases underly general transfer rules

[2] we collected 10 dialogues of appointment scheduling between German and Japanese speakers that were translated by an interpreter. The data includes 566 noun tokens.

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.

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 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. 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)) and *kondo no kaigi – das nächste Treffen/die nächsten Treffen* (the next meeting(s)).

2.3 Preference Rules and the Default

[SQ93] presented a 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[3] and have to be translated with

[3] Such a stack is also used for different purposes, as for example zero pronoun resolution.

definite determiner. An option with weak preference is to copy the information on number from the previous mentioned entity.

In German sentences with copula predicates number agreement between subject and x-complement is preferred. This can be stated as a preference rule[4]. Consider the following examples:

getsuyoobi wa kyuujitsu desu – *Montag ist ein Feiertag* (Monday is a holiday)

and

getsuyoobi to kayoobi wa kyuujitsu desu – *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.

3 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. We have shown that combinating transfer and the search for information on number and definiteness reduces the problem to a reasonable extend. 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 encoded in a stack on pre-mentioned entities. 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.

[4] It is not a general rule, as the example *wir sind ein Projektteam* (we are a project team) shows.

References

[Als92] H. Alshawi. *The Core Language Engine*. Cambridge: The MIT Press, 1992.

[BOI94] 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.

[MN93] 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.

[SQ93] B. Schmitz and J. Quantz. Defaults in machine translation. KIT-Report 106, Technische Universität Berlin, 1993.