Translation Knowledge Recycling for Related Languages

Michael Paul

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, 619-0288 Kyoto
Japan
paul@slt.atr.co.jp

Abstract

An increasing interest in multi-lingual translation systems demands a reconsideration of the development costs of machine translation engines for language pairs. This paper proposes an approach that reuses the existing translation knowledge resources of high-quality translation engines for translation into different, but related languages. The lexical information of the target representation is utilized to generate the corresponding translation in the related language by using a transfer dictionary for the mapping of words and a set of heuristic rules for the mapping of structural information. Experiments using a Japanese-English translation engine for the generation of German translations show a minor decrease of up to 5% in the acceptability of the German output compared with the English translation of unseen Japanese input.

Keywords

reuse of knowledge resources, multi-lingual extensions, related languages

1 Introduction

One of the biggest problems for the development of high-quality, multi-lingual translation engines is the high cost of adapting the underlying translation algorithm to multiple language pairs. In particular, the lack of resources (dictionaries, bilingual corpora, etc.) for "uncommon" language pairs forms a bottleneck for multi-lingual extensions.

The basic idea of this paper, as described in Section 2, is to devote efforts to the development of translation engines between the main linguistically different languages and to reuse the translation knowledge of these systems for translation into languages closely related to the target language. These languages must have similar grammatical characteristics so that the linguistic information contained in the target representation, e.g. a parse tree, can be mapped to a corresponding representation for the related language and so that this information can be used to generate the translation output.

Our approach does not depend on any specific language or translation engine, but simply requires an internal target representation containing structural and word information. Section 3 describes the translation engine and the knowledge resources used for the evaluation of our approach. The results of generating German output based on the translation of Japanese spoken-language utterances into English are summarized in Section 4.

In this approach, the English parse tree is mapped to a corresponding German one by substituting word phrases according to an English-to-German transfer dictionary and applying heuristic rules defining grammatical equivalences in both languages. Some comments on the feasibility of our approach and future perspectives are given in Section 5.

2 Translation Knowledge Recycling

Our aim is to find an inexpensive way to provide multi-lingual extensions of existing translation engines. A simple way to achieve this goal is the concatenation of the respective engines by using a text-based interface as illustrated in Figure 1. In this system, the source language (SL) is translated into an intermediate language (IL) by the first translation engine (TE1), and this translated text becomes the input of the second engine (TE2) that generates the output in the target language (TL).

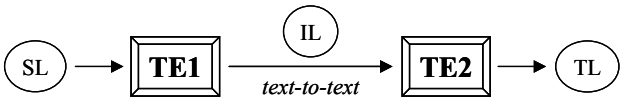

Figure 1: Text-to-text interface

The drawbacks of this scenario are the high costs of developing both translation engines as well as error magnification due to the isolated translation steps. However, if we could find a way to reprocess the translation knowledge of the intermediate language to directly generate the output in the target language, the costs would be drastically reduced.

The risk of such a recycling step is the lack of linguistic knowledge required in the target language. Therefore, the reuse of translation knowledge makes sense only for closely related languages that share similar grammatical characteristics.

2.1 Language representatives

According to their historical relatedness, languages can be grouped into so-called *language families* as illustrated in Figure 2. The families are marked in bold face. The languages on the same "branch" share certain features not shared by languages of other families.

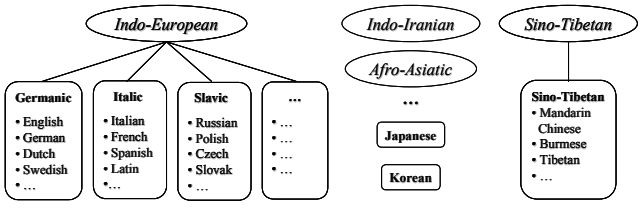

Figure 2: Language families

Based on the identification of similar characteristics between languages within the same language family, we might be able to find a representative for each family.

Examples for such representatives could be *English* for the Germanic languages, *Russian* for Slavic languages, or *Mandarin Chinese* for Sino-Tibetan languages. This would enable us to focus development efforts on high-quality translation engines between representative languages. Accordingly, we could concentrate on complex translation problems between completely different languages, whereas related languages could be dealt with in a more ad-hoc way. Moreover, there are several languages such as Japanese and Korean that do not belong to the same family but are grammatically similar, and thus potential candidates for the recycling of translation knowledge.

2.2 Mapping of translation knowledge

The basic units required for the generation of the translation output are words, their inflectional characteristics, and the features determining the grammatical context in which they occur.

On the word level, bilingual dictionaries can be used to define equivalent expressions. However, single words of the first language might correspond to more complex word phrases of the second language, and vice versa. Therefore, a word phrase dictionary is required to reduce word selection ambiguity caused by 1:1 word translation. Moreover, a morphological dictionary for the definition of the inflectional attributes of the target words is necessary for generating the translation output.

On the sentence structure level, the grammatical role of a specific word phrase is defined in the linguistic representation of the first language. Due to the relatedness of the languages, grammatical functions should be marked in a similar way in both languages. The identification of corresponding generation markers allows us to define rules mapping the linguistic knowledge from the first to the second representation. However, even if the grammatical functions are similar, the realization of the grammatical role during generation might differ between the languages, e.g., word order variations. This kind of language-dependent information has to be encoded in the generation process.

2.3 Recycling effect

Our approach of reusing existing translation knowledge leads reduced costs of multi-lingual extension, because we can limit the number of language pairs to language representatives.

Moreover, the costs of multiple full-scale translation engines can be reduced to those of developing a transfer dictionary and a generation dictionary, and these knowledge resources are already frequently available, at least for common languages like English.

The most difficult part of the translation process is carried out within the translation engine, e.g., a Japanese-to-English translation engine has to deal with problems like the recovery of the sentence subject, which is frequently omitted in Japanese but required in English (Yamamoto & Sumita, 1998). Similar to English, German is also a language that requires a subject. Thus we could benefit from the Japanese-to-English efforts by simply mapping and reusing the recovered subject for the generation of German translations.

Furthermore, the number of generation markers utilized in a specific language is limited. Therefore the compilation of mapping rules for related languages becomes inexpensive.

The disadvantage of knowledge resource recycling is the possible lack of translation knowledge required in the target language, i.e., grammatical information of the target language omitted or without any equivalence in the source language. How far this phenomenon limits the feasibility of our approach will be discussed in Section 4.3.

3 Framework

In Section 3.1 we give an overview of the translation engine. The knowledge resources and mapping algorithm of the proposed system are described in Section 3.2. Finally, an example of the reuse of translation knowledge is given in Section 3.3.

3.1 Translation Engine

The translation engine used for our experiments consists of a spoken-language machine translation system capable of bilingual translations between Japanese/English (JE). This transfer-driven translation system (TDMT) uses a *constituent boundary parsing* method (CBP) in an example-based framework. The input sentence is incrementally parsed by matching meaningful units of linguistic structure (patterns) with a chart-parsing algorithm. Given a set of translation examples, TDMT tries to find the "closest" examples to the structured input by using a *semantic distance calculation* (SDC) (Sumita et al., 1999).

By simulating the translation of the closest examples, the empirical transfer knowledge is applied to the source structure, resulting in a corresponding target structure, that can be used to generate the translation (cf. Figure 3).

3.2 Recycling system

The input of the proposed system (JeG) consists of the linguistic knowledge contained in the target representation of the JE system. The mapping algorithm for recycling the English translation knowledge is introduced in Section 3.2.3. First, it substitutes the English words in the English parse tree with corresponding German words by using a transfer dictionary (cf. Section 3.2.1). In the second step, the generation markers at each node of the parse tree are mapped by using a set of heuristic rules (cf. Section 3.2.2). The resulting German parse tree is then utilized to generate the translation output as described in Section 3.2.4.

3.2.1 Transfer Dictionary

The EG transfer dictionary for mapping English word compounds to corresponding German ones is created automatically from existing resources.

In order to reduce costs, we reused available Japanese-to-English (JE) and Japanese-to-German (JG) dictionaries created for the domain of our evaluation data by simply joining both dictionaries while using Japanese as the pivot language.

In general, any available EG dictionary could be used, but the joining of the JE (J=1 to E=n words) and the JG (J=1 to G=m words) dictionaries results in a word phrase dictionary for EG (E=n to G=m). Each entry consists of one

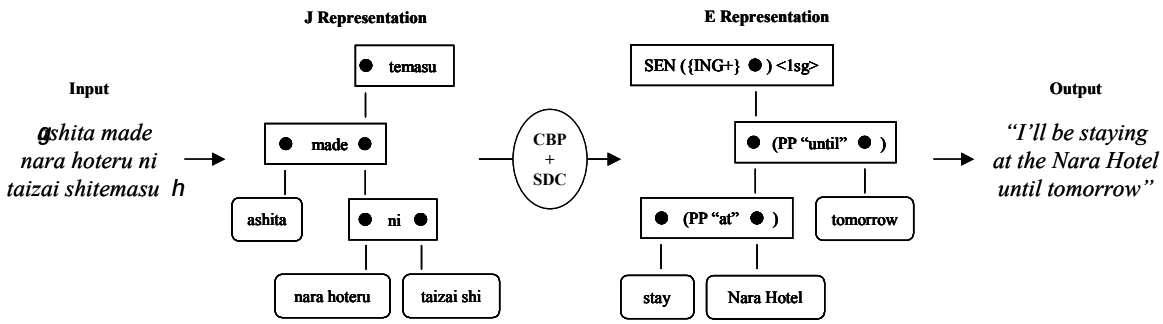

Figure 3: JE translation knowledge

or more part-of-speech tagged source words assigned to one or more target expressions as illustrated in **Fejl! Et bogmærke kan ikke henvise til sig selv.**.

1:1 (ADJ "German")
→ (ADJEKTIV "deutsch")
(CN "hotel")
→ (NOMEN "Hotel")

1:m (CN "vacancy")
→ (NP (ADJEKTIV "freies") (NOMEN "Zimmer"))
(V "hurry")
→ (VP (REFLEXIVPRONOMEN "sich") (VERB "beeilen"))

n:1 (ADJ "additional") (CN "charge")
→ (NOMEN "Aufpreis")
(CN "baggage") (CN "claim") (CN "area")
→ (NOMEN "Gepäckausgabe")

n:m (BEV "be") (ADJ "different")
→ (VP (ADJEKTIV "verschieden") (HILFSVERB "sein"))
(V "turn") (ADV "left")
→ (PP (PRAEP "nach") (ADVERB "links")) (VERB "abbiegen")

Figure 4: Word phrase dictionary

Additional costs for hand-checking automatically created dictionaries cannot be avoided, but research efforts are already under way to minimize these costs (Bond, 2001).

3.2.2 Mapping Rules

The heuristic mapping rules are defined by hand. First, we extracted all grammatical markers used in the JE training data and assigned German equivalents, e.g., the English direct object marker *OBJ* is mapped to the accusative complement marker *AKK-OBJ*, as illustrated in Figure 5.

In the second step, the created rules were verified by using a subset of the JE training data. One thousand utterances were translated by the JeG system, and the mapping rules were adjusted for translation errors in the context of the training sentences.

(OBJ $x)
→ (AKK-OBJ $x)
({ING+} $x)
→ $x
({PAST+} $x)
→ (ATTR $x :tense IMPERFEKT)
(SEN (RELP "which") $x)
→ (REL-S (RELATIVPRONOMEN "welches") $x))

Figure 5: Mapping rules

An investigation into the resulting rule set revealed the following rule clustering according to their functionality.

- *sentence structure*, e.g., type of subordinated sentences
- *phrasal structure*, e.g., word order within a phrase
- *inflectional marker*, e.g., number or tense
- *omission*, e.g., E markers without G equivalent

3.2.3 Mapping Algorithm

The mapping algorithm of the translation knowledge consists of two steps as described in **Fejl! Et bogmærke kan ikke henvise til sig selv.**.

E

Step 1: map word sequence

(1) *tree* ← parse-tree (translate_JE(*input*));
(2) *E-words* ← extract_words(*tree*);
(3) **until** *E*-words=Ø **do**
(4) (*E-phrase*, *G-phrase*)
← look_up_longest_match(*E-words*);
(5) tree ← substitute (*G-phrase*, *E-phrase*, *tree*);
(6) E-words ← remove (*E-phrase*, *E-words*);
(7) **end(until);**

Step 2: map generation marker

(8) **depth-first** (*tree*)
(9) **foreach** *node* **in** *tree* **do**
- */* LHS - left_hand_sideof rule */*
- */* RHS - right_hand_sideof rule */*

(10) *rule* ← match_LHS(*rules*, *node*);
(11) tree ← substitute (RHS, LHS, *tree*);
(12) **end(foreach);**
(13) **end(depth-first);**
(14) **return** *tree*;

Figure 6: Mapping algorithm

First, the source words in the parse tree were replaced with corresponding target ones according to the word phrase dictionary. The sequence of words contained in the nodes was extracted from the parse tree. If this sequence could be matched in the dictionary, the respective source words were replaced with the corresponding target words. Otherwise, the word sequence was reduced from right to left by one word and the dictionary look-up was repeated until a match was found.

In the second step, the parse tree was traversed depth-first, substituting source with target generation markers according to the defined mapping rules. The left-hand side of each mapping rule was applied at each node and in the case of a match the substructure is modified according to the right-hand side of the selected mapping rule.

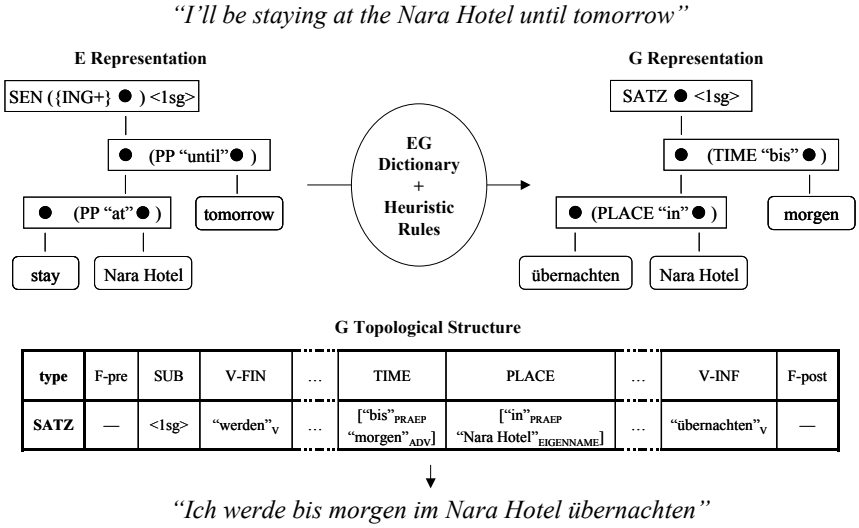

Figure 7: Recycling example

3.2.4 Generation

In contrast to more configurational languages like English, languages with partially free word order like German im-pose some additional burden on the generation of lin-guistic knowledge contained in the target representation.

We utilize an approach for clause syntax of the target language by employing the notion of *topological fields*, whereby sentence patterns are described as combinations of structural units. The linearization of these fields determines the clausal word order within the respective sentence pattern. The constituents of the target representation are updated to these fields according to their grammatical role marked in the mapped parse tree (Paul et al., 1998).

In order to generate the mapped German translation knowledge and to take into account word order variations of German, we only had to extend the topological field definitions used for English.

On the word level, we used a morphological dictionary, automatically extracted from the CELEX database (Piepenbrock, 1995), to generate the surface words based on the grammatical context of the respective phrases.

3.3 Recycling Example

An example of the mapping of an English parse tree and the generation of the corresponding German translation is given in Figure 7.

Some generation markers, like the subject marker *<1sg>* (first person, singular), are simply passed along without any modification. Others, like the English inflection marker *{ING+}* (progressive form), do not have any equivalence in German, and thus are omitted. The majority of rules, however, assign markers for corresponding grammatical functions, e.g., the prepositional phrase *(PP "until" ...)* is converted to a temporal expression *TIME* and *(PP "at" ...)* is mapped as a locative expression *PLACE* in order to take into account word order variations of the underlying topological structure of the German target representation.

4 Evaluation

In order to prove the feasibility of our approach, we applied our system to the same data set (cf. Section 4.1) using the same criteria (cf. Section 4.2) as for the evaluation of the JE system (Sumita et al., 1999).

4.1 Evaluation data

The data used for our experiments consist of the ATR-ITL Speech and Language Database containing Japanese-English spoken-language dialogs in the travel domain (Takezawa, 1999).

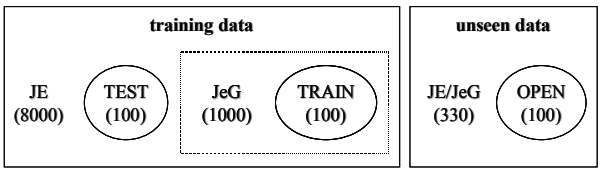

Figure 8: Evaluation data sets

To evaluate our approach we utilize three different data sets as illustrated in Figure 8. Each set consists of 100 randomly selected utterances. The TRAIN data was used for the training of the JE translation engine and the JeG system. In contrast, the TEST set consists of JE training utterances unseen by the JeG system. Finally, the utterances of the OPEN set were used for the open evaluation of both systems.

4.2 Evaluation criteria

The JeG system was applied to all three data sets and the translation results were evaluated by two German natives with knowledge of Japanese based on the same guidelines used to evaluate the JE system.

First, the evaluators read only the translation and retained the information they gathered. Then they referred to the Japanese input and identified the main information that has to be expressed in the translation. The extent to which the main information in the Japanese input corresponds to the translation output is evaluated based on the following four ranking options.

(A) *complete and accurate translation*: all of the main information is covered and expressed naturally, and the translation is immediately understandable.

(B) *fair translation*: the information is partially lacking or incorrect. There are some grammar mistakes or missing or misleading parts. However, the main information in the Japanese input can be easily obtained from the translation.

(C) *acceptable translation*: at first glance, it is difficult to obtain the information in the Japanese input from the output. However, based on the context, it is possible to reconstruct parts of the information from the input.

(D) *invalid translation*: the primary information is lacking, seriously incorrect, or one cannot make any sense out of what is being said.

4.3 Evaluation results

The results of our experiments are summarized in Table 1.

rank	TRAIN	TEST	OPEN	OPEN (JE)
A	76%	58%	56%	56%
B	15%	21%	18%	23%
C	9%	13%	13%	11%
D	0%	8%	13%	10%
A+B	91%	79%	74%	79%
A+B+C	100%	92%	87%	90%

Table 1: Evaluation of the JeG system

For the TRAIN set we achieved an acceptability of 100%, but the existence of 9% of rank C sentences shows that not all of the target phenomena could be covered accurately.

The TEST results show a large drop in accurate translations for correct English input sentences unseen by the system, but 79% are still at least fair and only 8% of the data were not acceptable.

Comparing the results of the OPEN test set, we see only a minor performance drop of up to 5% between the JE and JeG system, proving the feasibility of our recycling approach for related languages like English and German.

Furthermore, the implementation of the JeG system lasts only several months. Most of the time was spent on hand-checking the EG dictionary and verifying the heuristic rules. However, compared to the development costs for the translation engine, the results are quite promising.

5 Discussion

The evaluation results show that our approach still has to deal with translation problems like word disambiguation or structural differences even for related languages. However, the minor decrease in performance for open test data and low development costs demonstrates the feasibility of our approach to recycling translation knowledge for related languages.

The performance and coverage of our system depends on the utilized translation engine. However, the resources used for the mapping of translation knowledge are easy to extend and language-dependent target information, e.g. word inflection, can be handled with appropriate generation models for the target language. Therefore, upscaling the system to other domains should not lead to a tremendous increase in costs or decrease in the system performance.

In our experiments, we used English as the representative language for the Germanic language family, even if German would be a better choice due to its lexical richness. In that scenario, the omission of translation knowledge could be eased by mapping a richer representation to a poorer one. However, considering the available resources for other languages, English seems to be the most obvious choice.

Furthermore, we applied our recycling approach to the output language of our translation engine. However, we might also be able to apply our approach for languages related to the source language of an engine. Given a parser for the related language, we could map the internal representation to the source language and reuse our engine for the translation into the target language, e.g. from German to English to Japanese.

We also plan to apply our system to related languages outside the same language family, e.g. translation between Japanese and Italian through English.

6 References

Bond, F., Sulong, R.B., Yamazaki, T., and Ogura, K. (2001). Design and Construction of a machine-tractable Japanese-Malay Dictionary. In Proc. of the Machine Translation Summit VIII (to appear). Santiago de Compostella, Spain.

Paul, M, Sumita, E., and Iida, H. (1998). Field Structure and Generation in Transfer-Driven Machine Translation. In Proc. of the 4th Annual Meeting of the NLP (pp. 504--507). Fukuoka, Japan.

Piepenbrock, R. (1995). CELEX Lexical Database (Dutch, English, German), Version 2.5. Max Planck Institute of Psycholinguistics. Nijmegen, Netherlands.

Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, K., Ishikawa, K., and Shirai, S. (1999). Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX approach. In Proc. of the Machine Translation Summit VII (pp. 229--235). Singapore.

Takezawa, T. (1999). Building a bilingual travel conversation database for speech translation research. In Proc. of Oriental COCOSDA Workshop.

Yamamoto, K. and Sumita, E. (1998). Feasibility Study for ellipsis resolution in dialogues by machine-learning techniques. In Proc. of the 17th COLING (pp. 1428--1434). Montreal, Canada.