Semi-automatic Acquisition of Machine Translation Knowledge from Examples

Fuji REN
Institute of Information science ,
Hiroshima City University,
3-4-1, Ozuka-Higasi,Asa-Minami-Ku,Hiroshima,
731-31, Japan
ren@its.hiroshima-cu.ac.jp

Lixin Fan
Digital Express Corporation
3-601, Ozuka-Nisi 6-14, Asa-Minami-Ku, Hiroshima,
731-31, Japan

Jian-Yun NIE
DIRO, Université de Montréal
C.P. 6128, succ. centre-vile, Québec,
H3C 3J7 Canada
nie@iro.umontreal.ca

Youliang JIAN
Computer Institute,
Dalian University of Technology
116023, Dalian, China
newyou@dlut.edu.cn

Abstract

A crucial problem in rule-based machine translation is the acquisition of translation knowledge. Many studies have been conducted for automatic acquisition in the past, but they require a great deal of annotated examples. In this paper, we describe a semi-automatic acquisition from translation examples in Japanese-Chinese environment. Whenever necessary, the process interacts with a user (a linguist) who will provides additional information and constraints for generalizing the translation examples. The semi-automatic approach is more efficient in the sense that less examples are required than automatic approaches. Once the acquired knowledge is integrated with an MT system, we observe an improvement in translation quality.

Keywords: Machine Translation, Example, Translation Knowledge

1. Introduction

Many Machine Translation (MT) systems use rules to transfer a sentence from the source language to the target language. It is a pre-requisite that the rules used have a good coverage of the linguistic phenomena of the domain of interest. How to acquire the necessary translation rules is a crucial problem. In many MT systems built so far, translation knowledge is often edited and input manually. Beside the amount of tedious manual work, the manual editing may also produce many errors. In addition, it is difficult for the built system to deal with particular problems in a given application (for example, to translate legal documents). It is important to endow MT systems with the capacity of acquiring new knowledge during its utilization.

The problem of automatic knowledge acquisition in MT has been investigated in a number of studies (Furuse, 1992; Kaji. 1992; Kitamura, 1996; Sato, 1995; Watanabe, 1993). Knowledge is usually extracted from translation examples provided to the system. However, it is often assumed that lexical, syntactic and semantic analyses may be proceeded successfully in both source and target languages. This assumption cannot be made for languages such as Chinese because syntactic analysis is still problematic (Sun, 1994; Yu, 1995). If a completely automatic acquisition is applied to Chinese the errors produced during Chinese analysis would make the resulting rules almost useless. So, from a practical point of view, we think it is better to limit the requirement for automatic analysis of Chinese, and to ask the user for help when the system does not have all the necessary knowledge. This point of view agrees with (Kay, 1997) and several other studies (Langé et al. 1997; Dagan and Church, 1997; Zajac and Vanni, 1997; Ren et al. 1996).

Beside this, automatic approaches also have the following problems:

- An example is often given to illustrate the translation of certain word or linguistic phenomenon. It may not be exemplary for other words involved. An automatic acquisition process cannot make difference between them. The same process is applied uniformly to every word. Undesired rules (noise rules) may be extracted. One can always apply statistical criteria to extract translation rules that are verified by a good number of examples. It is often difficult to obtain sufficient examples.
- Extracting rules from examples also implies generalization. One can generalize examples to several different levels. For example, in some cases, the translation rule extracted may still apply if a word in the sentence is replaced by another one in the same syntactic category, or in the same semantic category. The appropriate level of generalization of an example cannot be recognized automatically. The only way to generalize in an automatic approach is to observe a number of similar examples. Some generalization may be made only if all the examples agree with it. We have the same problem of requiring a great deal of examples.

In our semi-automatic acquisition approach, we assume the presence of a "user" who may provide the system with additional information. A user is a person who has linguistic knowledge on both source and target languages. He will be asked to indicate the focus word of an example, to annotate the corresponding components in both sentences, and to choose the appropriate level of generalization for the example. The user works in interaction with the system. In this way, one does not need to annotate all the information uniformly for all the examples. If the system is able to produce a translation rule with the knowledge already acquired, the user is only asked to confirm it. Unnecessary manual annotation may be saved.

Semi-automatic acquisition has been investigated in several other studies. For example (Dagan and Church, 1997) uses a semi-automatic approach to extract terminology translation glossary. In our case, more structure is concerned: we deal with not only word translation, but also sentence structure.

Our acquisition process is composed of the following three steps:

- automatic Japanese input analysis
- user-assisted annotation of the example
- construction of a translation rule

The acquisition process described in this paper has been integrated into a Japanese-Chinese MT system, JC-TRANS (Ren 1996; Jian 1997). The system possesses an initial set of knowledge. It is then augmented with the knowledge acquired semi-automatically. Our preliminary experiments show that with the new knowledge, the quality of translation is greatly improved, and the extraction of rules prove to be much more efficient than the manual process. In the remaining of the paper, we will describe the three steps in more detail.

2. Input analysis

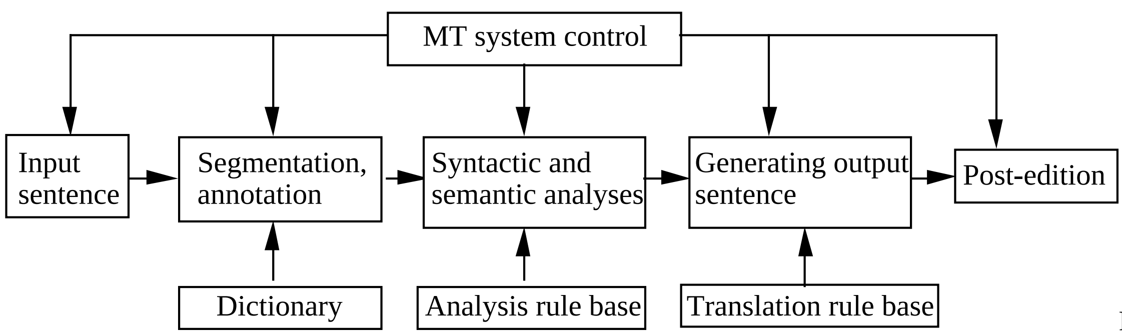
Both word segmentation and sentence analysis for Japanese are part of the translation system JC-TRANS. This system is shown in Figure 1.

The acquisition process takes advantage of the existing segmentation, annotation and analysis modules of JC-TRANS. In our present study, we are only concerned with Japanese to Chinese translation. Given a Japanese sentence or phrase, it is first segmented into words. Then syntactic and

semantic information is retrieved from the dictionary and associated to each word. There are two types of words in Japanese (and Chinese): "independent words" and "associated words". The first type designates the words which are meaningful alone and may be used independently in sentences. For example, verbs, nouns. On the other hand, associated words are not meaningful alone. They should be used in association with independent words in sentences. We call an independent word, together with its associated word, a "segment". Below is an example of the segmentation result:
Input:â~Ç•ÉhÉãÇ…î°Ç•ÅB
(change Japanese yens to American dollars)
Segmentation result:
â~.Ç•/ÉhÉã.Ç…/î°Ç•
where '/' is the symbol separating different segments and '.' separates an independent word and its associated word.

In most cases, the syntactic and semantic analyses of the Japanese sentence for details) allows us to recognize the following information for each independent word (Ind_Word) in the sentence:

- Syntactic category (Syn_Cat)
- Semantic category (Sem_Cat)
- Case
- Associated word (Ass_Word).



Figur
e 1. Structure of JC-TRANS system

Ind_Word	**Syn_Cat**	**Sem_Cat**	**Case**	**Ass_Word**
â~	NOUN	money	OBJ	Ç•
ÉhÉã	NOUN	money	D_OBJ	Ç…
î°Ç•	VERB	abstract transfer	VERB	

Table 1. Result of the analysis

In JC-TRANS, a set of semantic categories has been defined. For the moment, it is not sufficiently fine-grained, especially for verbs. However, we are able to validate our approach with the limited number of semantic categories. Sometimes, the user's help is required during the analysis in order to solve ambiguities and to add additional information. The result of the analysis for the above sentence is shown in Table 1 (where D_OBJ means a direction object).

The Chinese translation is also submitted to JC-TRANS. Although it is more difficult to analyze Chinese than Japanese, with the help of the user and the information obtained from the Japanese input, we can obtain a correct analysis result.

We notice that many examples contain, beside the words that affect the translation, other words that do not. For example, If we insert "êÁ" (thousand) in the above example as follows:

êÁâ~Ç•ÉhÉãÇ…í°Ç•Å

(change a thousand Japanese yens to American dollars)

the translation of "í°Ç•" (change) is not affected at all. We remove these words from the input in order to create a cleaner structure. A word or a segment which plays a modifier role in the input, which is not the focus word, is removed. From the cleaned input, a skeleton structure may be produced for the input by replacing each independent word by its case name. The principle of skeleton construction is similar to (Langé et al. 1997), but our method go further.

For the above example, we obtain the following skeleton structure:

Skeleton: OBJ . Ç• / D_OBJ . Ç… / VERB
OBJ: Word = â~,
Syn_Cat = NOUN,
Sem_Cat = "money"
D_OBJ: Word = ÉhÉã,
Syn_Cat = NOUN,
Sem_Cat = "money"
VERB: Word = í°Ç•,
Syn_Cat = VERB,
Sem_Cat="abstract transfer"

The first line describes the structure of the input, the other lines give specification of the independent words.

3. Example annotation

The user's manual annotation aims to indicate the following three types of information: the focus word of the example, the correspondence between independent words in the Japanese sentence and the Chinese sentence, and the constraints on the translation.

1) Focus word: Each example is presented to the system as an example for the translation of one particular word. In our present study, a focus word is a Japanese word. The rule extracted from this example will be attached to the focus word.

2) Correspondence between independent words: The user is asked to indicate, for each independent word in the Japanese sentence, its corresponding word in the Chinese sentence (if the correspondence is not found by the system). In so doing, we are able to drive the syntactic and semantic information for the Chinese sentence from the Japanese sentence.

3) Constraints: By constraints, we mean all the conditions required for applying the rule extracted from the current example. Constraints may also be viewed as those for the generalization of the example. They are set in order to avoid over-generalization. Constraints may be set at any of the following levels for every independent word in the skeleton: word, syntactic category, and semantic category. If a constraint is set at the word level, it means that the extracted rule may only be applied with the same word at that position. In the two other cases, the word may be replaced by another word (or phrase) of the same category. Therefore, a generalization of the example is made to some extent.

Ind_Word	Word	Syn_Cat	Sem_Cat
OBJ	â~	NOUN	money
D_OBJ	ÉhÉã	NOUN	money
VERB	î°Ç•	VERB	abstract transfer

Table 2. Possible constraints on independent words

Let us take the same example to illustrate the annotation process. We assume that the focus word in this example is "î°Ç•". The system will first find word correspondence as much as possible by using the existing knowledge. In case of incomplete or erroneous result, the user is asked to complete or correct it. Figure 2 shows the annotation result for word correspondence.

To set the constraints to this example, the user is asked to choose a level of constraint for each independent word among all the possible options show in Table 2.

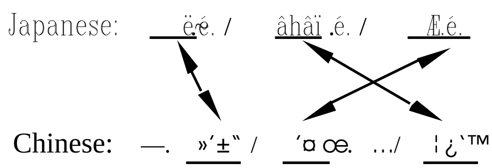


Figure 2. Annotated word correspondence

Suppose that the user chooses the constraints at the semantic levels for OBJ and D_OBJ, but the word level for VERB, then the rule extracted will mean: if OBJ and D_OBJ are in the semantic category "money", then a Japanese sentence of the form "OBJ Ç• D_OBJÇ…î°Ç• " may be translated into a Chinese sentence of the form "•— OBJ •‟ªª•… D_OBJ" in which, of course, OBJ and D_OBJ have been translated into Chinese by some other rules (or the dictionary).

In some (much less frequent) cases, a condition may also be set on the output. This is the case when the structure changes radically in the translation. For example, the Japanese noun phrase …Áª·•¯»Î (a NOUN NOUN structure) is translated to •¯»Î…Áª· (a VERB NOUN structure). In this case, we need to detect the change of syntactic category and of structure.

4. Construction of rules

To obtain a translation rule, we first construct a skeleton structure for the Chinese sentence. This is similar to the skeleton construction for the Japanese input. The skeleton construction makes use of the syntactic and semantic information associated to the Chinese output. This information is partly derived from the Japanese input due to the word correspondence annotated by the user, partly indicated by the user. The skeleton structure for our Chinese output example is as follows:

Skeleton: •— . OBJ / VERB . •… / D_OBJ
OBJ: Word = »´±‟ ,
Syn_Cat = NOUN,
Sem_Cat = "money"
VERB: Word = ´¤•œ,
Syn_Cat = VERB,

Sem_Cat="abstract transfer"
D_OBJ: Word = •¿`™,
Syn_Cat = NOUN,
Sem_Cat = "money"

According to the constraints the user annotated, we can construct a rule as follows:
- Output structure: The output structure is identical to the skeleton structure of the Chinese example.
- Specification of each variable in the skeleton: each variable may be attached with some condition (Cond) and/or with a translation (Trans). The condition for a variable is that chosen by the user. In addition, some information obtained (e.g. associated word) from Japanese analysis is also attached. A condition is to be satisfied by the component corresponding to the case in the Japanese input. If a translation is attached to a variable, it means that this is the only possible translation for the word in this situation.

For example, from the pervious treatments, we can construct a rule for our translation example which may be represented as follows:

î°Ç•: OUTPUT: •— .OBJ / VERB.•… /
D_OBJ
OBJ: Cond = (Sem_Cat = "money",
Ass_Word = "Ç• ")
#VERB: Trans = (Word = "•‶ ªª")
D_OBJ: Cond = (Sem_Cat = "money",
Ass_Word = "Ç…")

where # marks the corresponding position of the focus word. The rule means: if in a Japanese input, the verb "î°Ç• " has two words of the semantic category "money" as its "OBJ" and "D_OBJ", and they have "Ç• " and "Ç…" as their associated words, respectively, then the output is of the structure "•— . OBJ / VERB . •… / D_OBJ " where VERB should be translated as "•‶ ªª", and OBJ and D_OBJ are translated separately.

Notice that we only require that the Japanese input to have the same case structure (in stead of the same word order) as the example. This is because one may change the place of the OBJ segment and D_OBJ segment in Japanese.

This is the simplified form of translation rule which does not represent a radical change in structure (this is true in most cases). In case the structure is radically changed, we produce a more complex rule.

An extracted rule may concern the following aspects:
- the choice of the translation word
- the structure of a phrase or a sentence

From the example given above, one can easily see how rules can determine the structure of the output. Let us now describe some particular phenomena that the rules are able to capture.

A word in one language often corresponds to several different words in another language. For example, the Japanese focus word "î°Ç•" given in our example may be translated into –••ƒ,÷Œ¡•, ʹ¤•œ and •‶ ªª in Chinese. These words have different meanings, except the two last ones. It is important to choose the correct word translation. From the above example, we can determine that "î°Ç•" in Japanese should be translated into "•‶ ªª" (change) in Chinese in the specified context (where the objects are "money"). If in other examples in similar situations, the word "î°Ç•" is translated to "ʹ¤•œ", then the system may conclude that both translations are possible.

For the following sentence, we have a different translation of the word "í°Ç•":

Japanese:Å@ò_ï•Ç•Å@/í°Ç•

Chinese: –••ƒ / ¬¤Œƒ

(modify an article / a thesis)

The extracted rule from this example is as follows:

í°Ç•: OUTPUT: VERB / OBJ
    #VERB: Trans = (Word = "–••ƒ")
    OBJ: Cond = (Sem_Cat = "writing",
                    Ass_Word = "Ç•")

The difference between the two rules lies in the semantic category of the object.

We also deal with other grammatical constructs such as adjectives and verbs. However, for the sake of space limitation, we do not give more detail for them.

5. Preliminary experiments

The acquisition process has been integrated with the MT system JC-TRANS. It has a dictionary containing 65 000 general words. In addition, there are five specialized dictionaries. The system contains about 7800 rules. Figure 3 shows the window of the JC-TRANS.

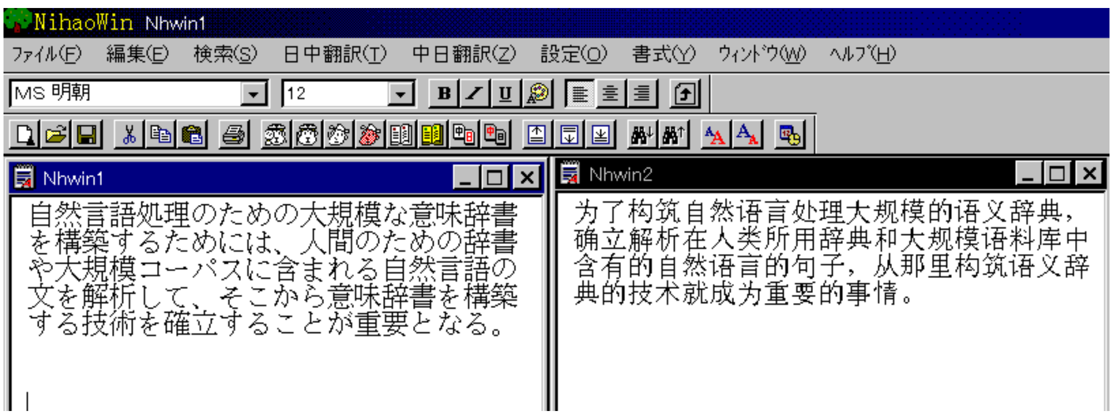


Figure 3. The window of JC-TRANS

A set of parallel sentences is selected as examples for the acquisition process. The examples come from bilingual dictionaries (examples in the dictionary items) as well as a bilingual book for Japanese learning. We gathered 3299 translation examples on 1600 focus words that be selected from a parallel corpora.. The sentences are annotated manually in interaction with the system. We used another set of 1582 sentences for the test. The longest sentence contains 59 words. We obtained 1503 correct translations. The success rate is 95%. This rate is high in comparison with the rules we edited manually a few years ago. We spent much time to edit 6800 rules, and the success rate using the rules for the same set of test sentences is only 79.6%. We can see that using the semi-automatic acquisition approach, rules may be acquired efficiently.

Most of the errors observed in the test are due to the incorrect analysis of modifier-modified relation for verbs. This problem needs to be further studied.

To illustrate the impact of the new rules on the translation, let us give the following example. Before the new rules being integrated, JC-TRANS translates the following sentence

ìØÇ•ÇǂÇÃÇǂÇ†ÇÍÇŒÇ•ÇÒÇ•ÇÒà·Ç§ÇǂÇÃÇǂÇ†ÇÈ

into

‶ •»Á••″–‶ª— •‶ ,À°µƒ£¨‶ •″–»´»ª•ª‶ª— •µƒ‶ ,À°µƒ ‶

A possible translation of this sentence is "also if there are [things] of the same meaning, also there are [things] of different meanings". This is not natural. We provided the following example about the word "Ç†ÇÈ" to the system:

Japanese:ñ{ÇǂÇ†ÇÍÇŒêVï•ÇǂÇ†ÇÈ

Chinese: °»″– È£¨‶ •″–±®÷•°£

(There are books, as well as journals.)

The system succeeded to extract a rule on how to translate sentences involving "Ç†ÇÈ". Once the extracted rules are added, JC-TRANS was able to translate the earlier sentence as follows:

°»″–‶ª— •‶ ,À°µƒ£¨‶ •″–»´»ª•ª‶ª— •µƒ‶ ,À°µƒ°£

(There are [things] of the same meaning, as well as [things] of different meanings.)

This second translation is of higher quality than the first one. In some other cases, although with the new knowledge, the translation is not completely natural, it is greatly improved. °

6. Conclusion

Although the acquisition process is at its first stage, it provides encouraging results. We believe that the acquisition process represents an effective way for the system to accumulate translation knowledge during the utilization. One of the advantages we can see is its flexibility: the knowledge extracted depends on the examples the user presents to it. So it may be applied in different area and produce different styles of sentences. For example, it is possible to extract a set of rules for legal document translation which is different from the rules for everyday language translation.

The study is far from being completed. There are a number of problems to be dealt with:

- Although the system may perform some automatic processing on the input sentence, the amount of the manual work is still high. In particular, if the user does not have much linguistic knowledge, the approach is unfeasible. So one of our future works will be extending the approach to extract rules from less controlled examples. For instance, we may use parallel corpora as a set of examples. This approach

has been used in a number of studies (Fung and McKeown, 1997, Dagan and Church, 1997), however, not for Japanese-Chinese. A difficult problem is to align words correctly. For Japanese-Chinese, we notice the particularity that some common characters are used in both languages. So they may be useful additional clues for aligning words.
- The kinds of knowledge we can extract now are relatively limited. We intend to enlarge the kinds of knowledge in our extraction. For example, the knowledge about associated words.
- When a great number of rules have been extracted, it is important to have means to deal with contradiction, redundancy and priority among different rules. We have integrated a few of such functions, for instance, grouping similar rules together. However, more investigation is needed.

Acknowledgments

This work has been partly supported by the Education Ministry of Japan under Joint Research Grant in Aid for International Scientific Research JCE-TC (09044179) and for Exploratory Research T+ Grammar (08878050), and a science grant (No. 69485002) from National Natural Science Foundation of China.

References

Dagan, I., Church, K. (1997) *Termight: Coordinating humans and machine in bilingual terminology acquisition*, Machine Translation, 12/1-2, pp. 89-107.
Fung, P., McKeown, K. (1997) *A technical word- and term-translation aid using noisy parallel corpora across language groups*, Machine Translation, 12/1-2, pp. 53-87.
Furuse, O. and Iida, H. (1992) *An Example-based Method for Transfer-Driven Machine Translation*, Proc. TMI-92, pp. 139-150.
Jian Y., Chang H., Wang X., Huang D. (1997) *A Japanese parser using dependence relation*, Language Engineering, pp. 258-265.
Kaji, H., Kida, Y. and Morimoto, Y. (1992) *Learning Translation Templates from Bilingual Text*, Proc. COLING-92, pp. 672-678.
Kay, M. (1997) *The proper place of men and machine in language translation*, Machine Translation, 12/1-2, pp. 3-23.
Kitamura, M. and Matsumoto, Y. (1996) *Automatic Acquisition of Translation Rules from Parallel Corpora*, Transactions of IPSJ, 37 / 6, pp. 1030-1040.
Langé, J.-M., Gaussier, É., Daille, B. (1997) *Bricks and skeletons: Some ideas for the near future of MAHT*, Machine Translation, 12/1-2, pp. 39-51.
Ren F. (1996), *SWKJC Machine Translation System Based on Translation Rules Acquired from Corpora*, Technique Report of Hiroshima City University,Technique Report of Hiroshima City University,HCU-IS-96-004, pp.1-31.
Sato, S. (1995) *MBT2: A Method for Combining Fragments of Examples in Example-based Translation*, Artificial Intelligence, 75, pp. 31-49.
Sun, M. (1994) *Chinese syntactic analysis using a multiple pass, deterministic algorithm and its application in text understanding*, Proceedings of Processing of Linguistic Information, pp. 103-168
Watanabe, H. (1993) *A Method for Extracting Translation Patterns from Translation Examples*, Proc. TMI-93, pp. 292-301.
Yu, S. (1995) *A hypothesis about restricted regular Chinese*, Proceedings of Chinese modernization, pp. 193-205

Zajac, R., Vanni, M. (1997) *Glossary-based MT engines in a multilingual analyst's workstation architecture*, Machine Translation, 12/1-2, pp. 131-151.