Proceedings of the 7th Machine Translation Summit (MT Summit VII), pp.496–500, 1999.

Improvement of Translation Quality of English Newspaper Headlines by Automatic Preediting

Takehiko Yoshimi
Software Business Development Center
SHARP Corporation

Ichiko Sata
Software Business Development Center
SHARP Corporation

Abstract

Since the headlines of English news articles have a characteristic style, different from the styles which prevail in ordinary sentences, it is difficult for MT systems to generate high quality translations for headlines. We try to solve this problem by adding to an existing system a preediting module which rewrites the headlines to ordinary expressions. Rewriting of headlines makes it possible to generate better translations which would not otherwise be generated, with little or no changes to the existing parts of the system. Focusing on the absence of a form of the verb of 'be', we have described rewriting rules for putting properly the verb 'be' into the headlines.

1 Introduction

As Japanese people have more opportunities to see English newspaper articles through WWW, it becomes more important to translate correctly them into Japanese. The most essential information of articles is expressed by the headlines. They have characteristic styles different from those of ordinary sentences, so as to convey much information with as concise expression as possible. The characteristic styles prevent a syntactic parser from generating appropriate parse trees for the headlines, thus decrease the quality of translations.

There are at least two possible solutions for this kind of problem. One is to extend syntactic rules so that the parser may analyse characteristic expressions. The other is to add to the existing MT system a preediting module which rewrites characteristic expressions to ordinary ones. Possible problems of the former approach include the difficulty of keeping the consistency and the portability of the syntactic rules. The latter approach (Shirai et al., 1993; Kim and Ehara, 1994) is preferable from the viewpoint of system design and maintenance.

Adopting the latter approach, we propose a method of improving the quality of translations for the headlines. In this paper we focus on a conspicuous phenomenon in the headlines, the absence of an appropriate form of the verb 'be', and formulate rewriting rules for inserting the verb 'be' in its proper place in the headlines, based on information obtained by morpholexical and rough syntactic analysis [1] . Rewriting of the headlines makes it possible to generate better translations, with little or no changes of the existing parts of the systems. While most systems would not probably accept, for example, the headline "Sales up sharply in June", they may be able to generate a satisfactory translation of the expression "Sales are up sharply in June" where "are" has been inserted. We have incorporated the proposed method into our English-to-Japanese MT system Power E/J, and carried out an experiment with 312 headlines as unknown data. Our method has satisfactorily marked 81.2% recall and 92.0% precision.

2 Preediting Module

2.1 Framework of Automatic Preediting

Figure 1 shows the flow of analysis in our experimental system with the preediting module. After the completion of morpholexical analysis, our preediting module runs to rewrite the original expression. Syntactic analysis rules is then applied to produce parse trees for the rewritten expression. If the initial syntactic analysis fails [2] , the process returns to the preediting module. In a second preediting phase, the module restrains some rewriting rules which were applied in the

[1] The rough syntactic analysis means a process using the procedures which will be mentioned in Section 3.2.

[2] In this paper, the failure of analysis means that no parse tree which covers the whole input was generated.

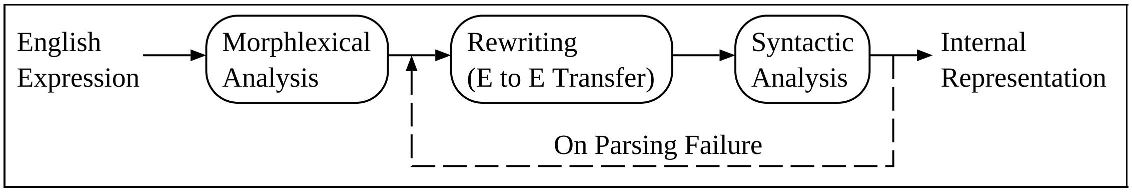

Figure 1: Flow of Analysis

first phase, and/or newly uses rules which were not applied, according to the certainty factor (See Section 2.2) given to the rules.

The preediting module examines from left to right on the list of morpholexical units whether a part of the list and the condition of the rewriting rules are matched, and it rewrites the parts where matches are established. Note that the module is not designed for the exclusive use of the headlines, but is a general framework which deals with ordinary expressions.

2.2 Form of Rewriting Rules

A rewriting rule consists of the rule number, the condition part, the action part, and the control instructions. The control instructions are the certainty factor and a set of the numbers of rules which are preferred to the rule.

Whether a part of the English input meets a condition is judged by procedures which examine morpholexical and syntactic features of the part.

Rewriting actions are classified into addition, deletion and substitution of English expressions, and insertion of preediting symbols proper to our experimental system. 54 preediting symbols are available in our system, including one for disambiguation of the word classes and one for disambiguation of the scope of a phrase and a clause. Disambiguation by inserting these symbols improves the efficiency and the quality of analysis.

A set of preferred rules given to a rule R is a metacondition about other rules which restrains the application of R: R is not applied in case that at least one of the rules whose number is a member of the set has already been applied. The set given to R may contains the numbers not only of the rules whose target overlaps with that of R, but also of the ones whose target does not.

Some rules are so reliable that their application probably improves the quality of translation, others are unreliable. Taking this into account, we introduce the certainty factor in order to control bad influences of less reliable rules upon quality. As the certainty factor, either “A”, “B” or “C” is given to each rule according to its reliability. A rule with “A” is applied before the initial syntactic analysis, and its effect remains even if the analysis fails and the process enters the second phase. We give a rule the certainty factor “A” where we empirically know that syntactic analysis probably ends in failure without application of the rule, and expect that the quality of translation generated from the rewritten strings, even if the analysis fails, would be higher than that from the original strings. A rule with “B” is applied before the initial syntactic analysis, and its effect is cancelled if the analysis fails. One with “C” is applied for the first time in the second phase.

3 Rules for Inserting Verb ‘be’

3.1 Preliminary Investigation of Headlines

Before describing rewriting rules for the omission of ‘be’, we investigated 284 headlines of news wire articles from Reuters (Lewis, 1997). The omissions of ‘be’ were found in 73 out of the 284 headlines. An expression which unites with ‘be’ to function as a finite predicate is here called a key. The keys which appeared in the 73 headlines can be classified into six types: past participles of transitive verbs, to-infinitives, present participles, predicative adjectives, prepositional phrases, and particles.

3.2 Matching Conditions of Rules

The matching conditions of the rewriting rules are mainly based on the following four characteristics of the headlines with the omission of ‘be’ which have been found in the preliminary investigation.

Existence of Noun Phrase in front of Key

In the headline where the verb ‘be’ is not expressed, there exists an NP in front of the key in question. An adverb may exist between the NP and the key as in the headline (H1).

(H1) Early gulf cash soybeans slightly firmer

We set the condition (C1).

(C1) An NP must exist either immediately in front of a candidate for the key in question, or immediately in front of the adverb which exists immediately in front of the candidate.

An NP which can meet the condition is recognised by the patterns:

$$\begin{aligned} \mathrm{NP} &= \mathrm{NP0}\ (\mathrm{PRE\ NP0})^{?} \\ \mathrm{NP0} &= (\mathrm{ADV}^{?}\ (\mathrm{ADJ}|\mathrm{Ven}|\mathrm{Ving}))^{?}\ \mathrm{NOUN}^{+} \end{aligned}$$

where the superscripts '?' and '+' mean repetition of once or less, and once or more respectively.

Nonexistence of Clauses Conflicting with Latent Clause

Changing the omission of 'be' to a visible form restores a finite predicate and makes it possible for a syntactic parser to recognise as a clause a part which was not recognised so by then. Here we call the part a latent clause. The subject of the latent clause is an NP which meets the condition (C1). Inserting "is" in front of "preparing" in the headline (H2) allows a parser to recognise "is preparing" as the restored finite predicate, and "Senate" as the subject.

(H2) Senate preparing for new U.S. budget battle

A decision on the insertion of 'be' can be made based on whether there is a clause which syntactically conflicts with the latent clause. While no clause syntactically conflicts with the latent clause in the headlines (H1) and (H2), there is such a clause in the headline (H3).

(H3) Reagan hopes to lift Japan sanctions soon

The latent clause in this headline, a clause whose subject is "Reagan hopes" and whose predicate is "are to lift", conflicts with the visible clause consisting of "Reagan" as its subject and "hopes" as its predicate. In cases like this, we heuristically select the interpretation as a visible clause.

Note that even if a visible clause exists in a headline, if no syntactic conflict occurs, we restore a finite form of 'be'. For example, the clause "trade row grows" in the headline (H4) does not conflict with the latent clause "U.S. official is to visit Japan", because they are separated by a conjunction "as" which indicates a clause boundary.

(H4) U.S. official to visit Japan as trade row grows

Based on these considerations we set the condition (C2).

(C2) There must not exist any clauses which conflict with a latent clause.

In the headline (H5), the latent clause "Three were sued ···" where "sued" is interpreted as a past participle, conflicts with the clause "Three sued ···" where it is interpreted as a finite form.

(H5) Three sued over ball valves for nine mile point

In case where a past participle whose spelling is identical to that of the finite form is a candidate for the key in question and a conflict occurs between the interpretations of the candidate, we impose the condition (C3) (See Section 3.2) instead of the condition (C2).

A clause boundary is specified in some cases by such a marker as a conjunction, a relative pronoun/adverb, and a comma; it is not specified in other cases. We deal with only the case that it is specified by a conjunction. Moreover we suppose that a headline consists of two clauses at most, and that the one is not the center-embedded clause of the other. Although syntactic analysis is required for making a strict examination of whether the condition (C2) is met, we use the following simple procedure for the examination.

(Step 1) If a conjunction serving as a clause boundary marker divides a headline in two parts, then send to (Step 2) one of the parts which includes a candidate for the key in question. If no conjunction is found, then send the whole headline to (Step 2).

(Step 2) Search the input string for a finite verb from left to right. If one is found, then search backwards for an NP whose head noun agrees with the verb in person and number [3] . If such an NP exists, then regard it as a subject of the finite verb and conclude the condition (C2) not to be met. Note that in case where a candidate for the key in question is a verb with ambiguity between the past participle and the finite form, skip the verb.

Condition on Past Participles

When a candidate for the key in question has ambiguity between the past participle and the finite form, a conflict occurs between the latent clause and the clause whose predicate is the finite form of the candidate. In cases like this, the condition (C2) prevents the appropriate insertion of the verb 'be' into the headlines.

To resolve the ambiguity, we examine whether an NP exists immediately behind the candidate and which verb patterns (Hornby, 1975) the candidate has. Interpreting the candidate as the finite means that the verb is in the active voice, whereas the other interpretation means that it is passive. If the candidate has neither SVOO nor SVOC as its verb pattern, and the object of the candidate exists, then the latter interpretation is syntactically impossible. Supposing that an NP which exists immediately behind the candidate becomes the object, we insert the verb 'be' if such an NP is not found.

If the candidate has either SVOO or SVOC pattern, an interpretation as passive is possible even if an NP exists immediately behind the candidate. To ensure the possibility, it is necessary to examine not only

[3] The search is conducted by using the same patterns that used for the condition (C1).

whether an NP exists immediately behind the candidate, but also whether the NP is followed by another NP. However, without making the closer examination, we inert the verb 'be' if the candidate has either SVOO or SVOC. One of the reasons is that there is a well-known heuristics that on the ambiguity between the past participle and the finite form, the former might give the correct interpretation in most cases (Uenoda, 1978).

Based on these considerations, we set the condition (C3).

(C3) When a candidate for the key in question has ambiguity between the past participle and the finite form, either the candidate must not be followed immediately by an NP, or it must have the verb pattern SVOO or SVOC.

According to this condition, the verb 'be' is properly inserted into the headline (H5) where no NP exists immediately behind "sued".

Nonexistence of Fixed Expression

We set the condition (C4) because it would be better to keeping a headline as it is in most of the cases where a candidate for the key in question and the preceding NP compose a fixed expression such as an idiom or a collocation.

(C4) A candidate for the key in question and the preceding NP must not compose a fixed expression.

The headline (H6) is not rewritten because the dictionary entry for "need" says that the word and to-infinitive can be regarded as a unit.

(H6) No need to state U.K. support for system — Lawson

3.3 Decision of Inflectional Form

To rewrite a headline properly, the preediting module should make a decision not only on whether it puts the verb 'be' immediately after the subject candidate or not, but also on the inflectional form of 'be' in the former case.

Normally the inflectional form must be decided based on grammatical information such as tense, aspect, and the person and number of the subject. However, we allow only present tense and make a distinction among "am", "are" and "is" according to the person and number of the subject. This is not so unnatural because headlines often express past events by present tenses (Shirai et al., 1997; Uenoda, 1978).

3.4 Control Instructions of Rules

Although the omission of 'be' can generally exist twice or more in a headline, the case was not found in the headlines subject to our investigation. Therefore we formulate rules so that an insertion of 'be' is made just for once: we give to a rule R a set of preferred rules, as mentioned in Section 2.2, which consists of the rule numbers of all rules except R.

The certainty factor we gave every rule is "B", which causes the cancellation of rewriting if the initial syntactic analysis fails and the process enters the second phase.

4 Experiment

Table 1 shows the results of experiments which were carried out on 284 headlines as known data and 312 headlines as unknown data. In evaluation, we regarded a rewriting as correct if both person and number of the inflectional form 'be' are correct even if its tense and aspect are not appropriate.

The causes of noises and leaks are shown in Table 2. One error in the known data and 6 errors in the unknown data should not be blamed on the rewriting rules but on the morpholexical analysis. Most of the errors results from the mis-judgement of the condition (C2). Furthermore the mis-judgements are caused by mis-recognition of the clause boundaries and mis-resolution of ambiguity of word classes. The latter problem can be almost solved by incorporating a method of identifying nouns and verbs (Takeda and Matsuo, 1993; Takeda et al., 1995).

Causes	Known		Unknown	
	Leaks	Noises	Leaks	Noises
Morph.	1	0	3	3
Cond. (C1)	1	0	0	0
Cond. (C2)	5	1	9	0
Cond. (C3)	0	0	0	1
Cond. (C4)	0	1	0	2
Other Conds.	1	0	4	0
Total	8	2	16	6

Table 2: Causes of Errors

Since all the rules are given the certainty factor "B", rewriting may be cancelled if the initial syntactic analysis fails. This has never happened in the experiment with the unknown 312 headlines as well as with the known 284 headlines.

5 Related Works

As far as we know, few studies have been made on the improvement of the quality of translation of English newspaper headlines by automatic preediting. Shirai et al. (Shirai et al., 1993) and Kim and Ehara (Kim and Ehara, 1994) have proposed frameworks of automatic preediting, but these studies do not involve

Keys	Known Data		Unknown Data	
	Recall	Precision	Recall	Precision
Past Part.	87.5%(21/24)	100%(21/21)	87.8%(36/41)	94.7%(36/38)
To-Inf.	100%(17/17)	100%(17/17)	88.2%(15/17)	88.2%(15/17)
Pres. Part.	91.7%(11/12)	100%(11/11)	62.5% (5/8)	100% (5/5)
Adj.	81.8% (9/11)	90.0% (9/10)	69.2% (9/13)	90.0% (9/10)
Prep. Phr.	83.3% (5/6)	83.3% (5/6)	66.7% (2/3)	66.7% (2/3)
Particle	66.7% (2/3)	100% (2/2)	66.7% (2/3)	100% (2/2)
Total	89.0%(65/73)	97.0%(65/67)	81.2%(69/85)	92.0%(69/75)

Table 1: Experimental Result

the technique of rewriting the headlines. Shirai et al. insert a rewriting module between the existing syntactic analysis module and the semantic analysis module. Kim and Ehara divide a long sentence into short ones to deal with the problem that syntactic analysis of a long sentence whose translation tends to fail because of its length.

As mentioned in Section 2, our framework of automatic preediting is not designed for the exclusive use of the headlines, but is a general framework. A comparison of our framework with the above-mentioned two can be made as follows. While in Shirai et al.'s system preediting is not activated until syntactic analysis is finished, our preediting module runs just after the completion of morpholexical analysis, which makes it helpful for parsing. Unlike Kim et al.'s method which splits a sentence into pieces (and supplies a subject for each piece if necessary), our rewriting actions consist of addition, deletion and substitution of English expressions as well as insertion of preediting symbols, such as a symbol for disambiguation of the word classes and one for disambiguation of the scope of a phrase and a clause.

6 Conclusion

Focusing on a phenomenon frequently observed in English newspaper headlines, the absence of a verb of 'be', we have formulated rewriting rules for putting properly the verb 'be' into the headlines, based on information obtained by morpholexical and rough syntactic analysis. In small experiments, our method has shown satisfactory results.

Further studies are necessary on 1) describing rules to rewrite the comma which, often serving as a coordinate conjunction, lowers the translation quality; 2) making use of information obtained from the main part of an news article in order to improve the accuracy of rewriting.

References

A. S. Hornby. (1975). "*Guide to Patterns and Usage in English*". Oxford University Press.

Y. Kim and T. Ehara. (1994). "An Automatic Sentence Breaking and Subject Supplement Method for J/E Machine Translation". *Trans. Inf. Proc. Soc. Japan*, 35(6):1018–1028. (in Japanese).

D. D. Lewis. (1997). "Reuters-21578 Text Categorization Test Collection, Distribution 1.0". http://www.research.att.com/~lewis/reuters21578.html.

S. Shirai, S. Ikehara, and T. Kawaoka. (1993). "Effects of Automatic Rewriting of Source Language within a Japanese to English MT System". In *Proc. TMI*, pages 226–239.

S. Shirai, Y. Oyama, Y. Nakao, M. Nishigaki, H. Ueda, and Y. Omi. (1997). "Characteristics of English Newspaper Article Headlines". Proc. 54th Annual Convention, 4B-1, Inf. Proc. Soc. Japan. (in Japanese).

M. Takeda and F. Matsuo. (1993). "A Method for Determining Verb of Sentences in Abstracts of Scientific and Technical Literature". *Trans. Inf. Proc. Soc. Japan*, 34(9):1931–1936. (in Japanese).

M. Takeda, J. Suda, N. Kusumoto, and F. Matsuo. (1995). "Identification of Nouns in Abstracts of Scientific and Technical Literature". *Trans. Inf. Proc. Soc. Japan*, 36(8):1828–1837. (in Japanese).

M. Uenoda and T. Fuse. (1978). "*How to Read English Newspapers*". Asahi Press. (in Japanese).