Technical Reports on Language Technology
Report No 34, 20018
http://www.njas.helsinki.fi/salama

Optimizing Rules in English to Finnish Machine Translation[1]

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Rule-based machine translation requires several sets of rules in various phases of procession. The analysis of source language requires rule sets for morphological disambiguation and syntactic mapping. In this report we exclude these rules and we concentrate on rules, which produce the correct translation in target language. Three types of rule files are needed, (a) rules for isolating MWEs. (b) rules for semantic disambiguation, and (c) rules for adding various types of tags, which will then be converted into surface form clitics and added to the word. Because the number of rules needed for translating unrestricted text tends to grow so that processing speed will be affected, it is important to find ways for downsizing the rule files. This paper will discuss some of such ways using the type (c) rules as examples.

Key Words: *machine translation, constraint grammar, mapping rules.*

1 Introduction

Translating unrestricted text from English to Finnish requires large lexicons and various types of rule files with tens of thousands of rules. The basic way of reducing rules is to use the default interpretation. For example, each word has a default meaning. This is selected, if no rule selects another interpretation. This requires that each word has the most obvious gloss first in the list, if the glosses are not weighted. The default interpretation can be extended further. The nominative case of nouns, adjectives, numbers and pronouns can be considered the default case. Therefore, nominative case need not be specifically marked; it is the default case. Also, in singular/plural dichotomy, singular can be considered as default. Only plural needs to be marked. In the discussion below, also other types of using default interpretations will be suggested.

The discussion below is based on the analysis of rule application of Finnish inflection rules in the corpus of 645,500 words of news text.

2 Commonly applied inflection rules

The most often applied mapping rule in the corpus was the rule, which adds the nominative tag to the subject of the clause (1).

---

(1)
```
MAP (@NOM) TARGET SUBJ (1 VFIN) (NOT *1 HAVE OR (HAVE) OR
("have_to") OR ("should") OR ("shall") OR ("must") OR ("ought") OR
("need") OR ("become") + (S-ELA) OR ("feel") OR ("lack") BARRIER
CLB OR PREP) (NOT *2 (AG-PART) BARRIER CLB) (NOT -1 (NUM-PL));
```

The rule adds the tag @NOM to the subject, but it has some restrictions to rule application. The rule applied 44,167 times in the corpus. If the nominative case of the subject would be defined as the default case, this rule could be omitted. Rules for the subject would be needed only for those cases, where the subject case is something else than nominative.

Also, the second commonest rule adds the tag @NOM, but this time to any noun, which is not mapped (2).

(2)
```
MAP (@NOM) TARGET N (NOT 0 MAPPED) (NOT 0 (%A>));
```

This rule applied 17,406 times in the corpus. This rule also could be omitted, if nominative interpretation would be defined as default.

A commonly applied rule is also the rule, which maps the accusative tag to the object. In Finnish, object has one of the three cases, genitive accusative, nominative accusative and partitive. The rule below (3) adds the tag @ACC to the object. This is the tag for genitive accusative, and, as the statistics show, it is the most common object case.

(3)
```
MAP (@ACC) TARGET OBJ (NOT 0 (INDEF) + PL) (*-1 (TRV) BARRIER CLBV
LINK NOT 0 HAVE LINK NOT -1 NEG);
```

Explanation: Map the tag @ACC to the object. The target should not be preceded by an indefinite article plus plural. On the left there should be a transitive verb (TRV), but it should not be the verb *have*, and the word next to the left should not be a negation word. Do not scan beyond the clause boundary or verb. The rule applied 13,609 times.

There is another rule for adding the nominative tag to the subject (4).

(4)
```
MAP (@NOM) TARGET SUBJ (NOT *1 NEG OR (S-ABL) OR (S-ADE) OR (AG-
PART) BARRIER CLB) (NOT 0 NUM);
```

Explanation: Add nominative tag @NOM to the subject. On the right there should not be a verb with a tag S-ABL or S-ADE or AG-PART. Do not scan beyond cpause boundary. The target should not be a number. The rule applied 12,348 times.

This rule also could be omitted, if nominative is defined as a default case for the subject.

The next in order is the rule for partitive case of the object (5).

(5)
```
MAP (@PAR) TARGET OBJ (*-1 (O-PAR) BARRIER CLBV OR PREP OR OBJ
LINK NOT 0 HAVE LINK NOT -1 ("so"));
```

Explanation: Add the tag @PAR to the object. On the left there should be a verb with the tag O-PAR. The verb should not be *have* and the word immediately to the left should not be *so*. Do not scan beyond the clause boundary or preposition or object. The rule applied 11,206 times.

The genitive structure is quite common in Finnish compounds. The rule (6) is restricted and applies only to clearly defined cases. Yet it applies often, because such structures are frequent in language.

(6)
```
MAP (@GEN) TARGET N OR PRON (-1 (M-GEN)) (NOT 1 N) (NOT -2
("number"));
```

Explanation: Add the genitive tag @GEN to the noun of pronoun, if the first word on the left has the tag (M-GEN). The next word to the right should not be a noun. The second word on the left should not be a number. The rule applied 10,019 times.

The next rule deals with locative, more precisely, inessive and adessive cases (7).

(7)
```
MAP (@INE) TARGET NPRON (*-1 (M-LOC1) BARRIER CLBV OR PREP OR
QUOTE LINK NOT 0 ("on")) (NOT 0 OUT OR SUBJ OR (WEEK)) (NOT 1
("ago"));
```

Explanation: Add the inessive tag @INE to a noun or pronoun. On the left there should be a preposition with the tag M-LOC1. This preposition should not be the word *on*. Do not scan beyond the clause boundary or verb or quotation mark. The target should not have the tag OUT or SUBJ or WEEK. The first word to the right should not be *ago*. The rule was applied 8,901 times.

Note that the tag M-LOC1 is a meta tag, referring to inessive and adessive. The selection between inessive and adessive takes place according to context conditions. In this case it is the tag OUT (outer locative). If the target does not have the tag OUT, inessive is selected. In the opposite case, adessive is selected.

The rule could also be implemented so that inessive is defined as default value for the tag M-LOC1. Then this rule would not be needed.

The next rule in order is still one for mapping the subject case (8).

(8)
```
MAP (@NOM) TARGET SUBJ (*1 HAVE BARRIER CLB LINK *1 EN BARRIER CLB
LINK NOT 0 HAVE) (NOT 1 PREP OR ("should") OR BE);
```

Explanation: Add the nominative tag @NOM to the subject. On the right there is the verb *have*, and after it somewhere to the right the tag EN, but this should not be the verb *have*.

In both cases do not scan beyond the clause boundary. The word after target should not be *should* or *be*.

This rule is for handling subject case in sentences with perfect or past perfect structure. The rule was applied 5,799 times.

The next rule deals with verb forms (9).

(9)
```
MAP (@EN-PERF) TARGET (EN) (-1 HAVE + (PRES) OR HAVE + (PAST) OR
HAVE + ING) (NOT -2 ("imagine") OR ("think"));
```

Explanation: Add the perfect tense tag @EN-PERF to the verb with the tag EN, if the first word on the left is the verb *have* plus present tense tag PRES or the verb *have* with the past tense tag PAST plus the gerund mark ING. The second word to the left should not be *imagine* or *think*.

The next rule deals with genitive (10). It is more general than the rule in (6).

(10)
```
MAP (@GEN) TARGET (%<P) (*-1 (M-GEN) BARRIER CLBV OR COMMA) (NOT -
1 (COMP));
```

Explanation: Add the genitive tag @GEN to the reading with the syntactic tag %<P. On the left should be a preposition with the tag M-GEN. Do not scan beyond clause boundary or verb or comma. The first word on the left should not have the tag COMP meaning that it is part in a compound noun. The rule applied 4,433 times.

In rule (6) the tag M-GEN must be immediately before the target. Rule (10) allows other words, such as adjectives, pronouns and numerals in between.

The next rule also deals with genitive (11).

(11)
```
MAP (@GEN) TARGET N OR PRON (-1 A OR NUM OR PRON OR (CAP) OR
(CAPALL)) (-2 (M-GEN)) (NOT 1 N) (NOT 0 SUBJ);
```

Explanation: Add the genitive tag @GEN to the noun or pronoun, if the first word on the left is adjective or numeral or pronoun, or a word with a capital initial letter (CAP or CAPALL). The second word to the left should be a preposition with the tag M-GEN. The first word to the right should not be a noun. The target should not be the subject. The rule applied 4,275 times.

The next rule deals with the second verb in a verb chain (12).

(12)
```
MAP (@3INF-ILL) TARGET (INF) OR (ING) (*-1 (V-3INF-ILL) BARRIER
CLBV LINK NOT -1 ("it")) (NOT -1 ("in order to") OR ("not_to") OR
M-MAPPED);
```

Explanation: Add to the infinitive verb the tag @3INF-ILL. On the left there should be a verb with the tag V-3INF-ILL, and the first word to the left from it should not be the pronoun *it*. Do not scan beyond clause boundary or verb. Immediately to the left from the

target there should not be the word cluster *in order to* or *not_to*, or any M-initial tag. The rule applied 3,584 times.

The need of this rule is questionable, because it requires that the preceding verb in the verb chain is marked with a certain inflection tag. This kind of inflection in Finnish is so common that there is motivation to consider it as default inflection in verb chain structures. Rules could be written only for those verbs, which deviate from the default inflection.

The next rule (13) deals with the case of the subject, which deviates from the default case, which is nominative.

(13)
```
MAP (@PAR) TARGET SUBJ (*1 BE + (NOGLOSS) BARRIER CLBV LINK 1 EN +
(CNT-ACT));
```

Explanation: Add to the subject the tag @PAR. On the right there should the verb be plus the tag NOGLOSS, and from it next to the right there should be a verb with the tag EN plus the tag CNT-ACT. The rule applied 3,533 times.

In these clause structures, the rule applies only to verbs, which have the tag CNT-ACT (continuous action). The rule is not fully safe, because the same verb may be sometimes used in the sense of continuous action and in other cases in the sense of finished action.

The next rule deals with the locative case called elative (14).

(14)
```
MAP (@ELA) TARGET NPRON (NOT 0 OUT OR SUBJ) (NOT -1 ("part_of"))
(*-1 (M-LOC2) BARRIER CLBV OR PREP);
```

Explanation: Add to the noun or pronoun the locative case tag @ELA. The target should not have the tag OUT or it should not be a subject. The next word to the left should not be *part_of*. On the left there should be a preposition with the tag M-LOC2. Do not scan beyond clause boundary or verb or preposition. The rule was applied 3,363 times.

Here again, the tag M-LOC2 stands for elative and ablative. The tag OUT is the criterion for choosing elative.

Here we could also use default definition. In that case we could rewrite the rule (14) as in (15).

(15)
```
MAP (@ELA) TARGET NPRON (NOT 0 SUBJ) (NOT -1 ("part_of")) (*-1 (M-
LOC2) BARRIER CLBV OR PREP);
```

And the ablative case could be chosen with the rule as in (16).

(16)
```
MAP (@ABL) TARGET NPRON (0 OUT) (NOT 0 SUBJ) (NOT -1 ("part_of"))
(*-1 (M-LOC2) BARRIER CLBV OR PREP);
```

The next rule deals with the genitive case of proper names (17).

(17)
```
MAP (@GEN) TARGET (CAP) + N (1 N) (NOT 0 ("river") OR ("lake"))
(NOT 2 N) (NOT -1 PREP);
```

Explanation: Add the genitive tag @GEN to the proper name. The next word should be noun. The target should not be the word *river* or *lake*. The second word to the right should not be noun. The preceding word should not be preposition. The rule was applied 3,264 times.

The next rule controls the use of locative illative/allative (18).

(18)
```
MAP (@ILL) TARGET NPRON OR (WH) (*-1 (M-LOC3) BARRIER CLBV OR PREP
LINK NOT 0 ("on")) (NOT 0 OUT OR SUBJ) (NOT 1 (PROPNAME) OR N)
(NOT *1 ("give") OR ("send") BARRIER CLB);
```

Explanation: Add the illative case tag @ILL to a noun or pronoun or question word. On the left there should be a prepositon with the tag M-LOC3, but this preposition should not be *on*. Do not scan beyond clause boundary or verb or preposition. The target should not have a tag, which belongs to the set OUT, and it should not be subject. The next word should not be proper name or noun. On the right there should not be the word *give* or *send*. Do not scan beyond clause boundary. The rule was applied 3,245 times.

3 Rarely applied rules

Above we have discussed such rules, which are frequently applied. The number of the rules in the file controlling Finnish inflection was about 1020. Most of the rules were applied only a few times, and many of them not even once. This happens, when the system is being developed and new solutions are tested. Some rules that were appropriate become obsolete, when some parameters change.

The occurrence of each rule was inspected, and it was possible to remove a total of 300 obsolete rules.

A word of warning is needed here. The rules were tested in a corpus of 645,500 words of news text. It is likely that some more rules would have applied, if the corpus were larger. However, because this rule file concerns more the sentence structure and less the lexicon, it is likely that not much harm will be caused if those rules are removed.
It is also possible that many of the now rarely applied rules could be merged with other rules by relaxing their constraints. However, here one must be very careful, because the danger of over-application is real.

4 Order of the rules

In the rule file discussed here, it is utmost important that the rules are in correct order. Once a rule has applied, no other rule coming later can apply to the reading. This is prevented, even if the later rule would be the correct one and the earlier rule the wrong one.

The guiding principle in rule ordering is to place the clearly defined rules first and the rules with scanning later. The guiding idea is: apply first those rules, which you can be sure about. Try then those rules, which are likely correct, but about which you cannot be sure.

5 Conclusion

In rule-based machine translation it is crucial that, in order to speed up processing, various methods for reducing rule writing are used. Default interpretation is one method, but also the careful control of the rule structure can speed up processing. In this paper we discussed the rule file for adding instructions for Finnish morphology in the English to Finnish translation system. It was found that some commonly applied rules could be removed by using default interpretation. It was also found that about 30 percent of the rules were obsolete when tested with the corpus of 645,55 words of news text.