Technical Reports in Language Technology
Report No 21, 20016
http://www.njas.helsinki.fi/salama

Integrating translation memory with rule-based translation system

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

Abstract

The integration of statistical translation with rule-mased translation is in many ways problematic. The basic problem with statistical translation is that it tries to guess rather than know. On the other hand, rule-based translation makes use of grammar and lexicon and produces grammatically correct language. It is not clear on which point of processing and how the statistical method should enhance the translation produced by the rule-based method. Even more difficult is to enhance the statistical translation with rule-based components. The use of translation memory to enhance rule-based translation is more secure to implement. When translation memory and rule-based translation are integrated, both work independent of each other. Some sentences are translated using translation memory, and others are translated using the rule-based system. One system does not try to improve the result produced by the other system. In fact, translation memory translates safely the sentences that are in the translation memory database. The rule-based system translates the rest.

Key Words: *machine translation, translation memory.*

1 Introduction

Translation memory (TM) is a collection of model translations, usually tranlated by humans. For practical reasons, the length of each piece of translation is a sentence. Also shorter word sequences can me included into the TM database, such as clauses. Shorter sequences have, however, two disadvantages. First, the sequence may be difficult to map within the rest of the sentence. Second, if part of sentence is translated using translation memory, the rest of sentence may be difficult to translate, because the linguistic information contained in the section translated by TM will be lost. Therefore the safest method for merging TM and rule-based MT is to handle only sentences in TM.

The rule-based translation system can translate all sentences. Why then include TM into the system? The reason is that language has a large number of expressions that recur in text. If such an expression is a sentence long, it is a good cndidate to be included into the TM. The sentences that are in the TM database are safe cases. They produce correct

translation and no correction is needed. Therefore, a large database of TM reduces the number of mistakes in translation and makes correction faster.

Discussion is going on in conferences and workshops on how the integration of TM and the translation system proper can be done. This report describes the solution, where some sentences are translated by TM and the rest of sentences are translated by the rule-based translation system.

2 Constructing the TM database

There are a number of ways to construct the translation memory database. The idea is that a sentence in source language (SL) has a corresponding sentence in target language (TL). In the subsequent implementation I use Beta rules for the purpose. In it, the string to be substituted, in this case the sentence in SL, terminates in semicolon ';' After this is one space, and after that is the translation in TL, which again terminates in semicolon ';'. In (1) is an example of such a rule.

(1)
Leo unakwenda wapi?; Where do you go today?;

The advantage of Beta rules compared with database columns is that usually translated model sentences do not need manual handling at all when converting into Beta rules. Necessary modifications, such as adding the semicolon, can be done with a script.

The accumulation of model translations is usually an ongoing process. The TM database, in our case the rule set, is constantly expanding, as soon as suitable model translations are at hand. Adding these new translations into the existing database can be done with a script, which also converts the sentences into Beta rules. For example, we have two more translations as in (2).

(2)
Utakuja kesho? Do you come tomorrow?
Nitakwenda sokoni. I will go to the market.

The script transforms them into Beta rules and adds them to the existing rule file (3)

(3)
Leo unakwenda wapi?; Where do you go today?;
Utakuja kesho?; Do you come tomorrow?;
Nitakwenda sokoni.; I will go to the market.;

New translations can be added individually or as a batch, whereby they must be saved into a separate file first.

3 Integrating the TM with Salama translator

The process goes in phases, so that first the text is put into sentence per line format. Then the text is run through the Beta rules to test whether the text has such sentences that are in

TM. If a sentence matches, it is immediately translated by the Beta rule. The rest of the text is left in its original form.

Then the text, which has translated sentences and untranslated sentences, runs through the translation routines of Salama. It checks each word, including the already translated sentences. Salama interprets English words as unknown tokens and copies them as such to TL.

Such sentences that were not translated by Beta rules, will be translated using the complex sequence of translation phases. In translation result it is hard to know which sentence was translated with Beta rules and which through the proper translation routines.

4 Controlling word forms that are found in both languages

It may happen that an English word is also a Swahili word, whereby the system tries to translate it. One example is a commonly occurring word 'do', which in Swahili means 'oh damn'. Such words must be 'injected', so that they will not be translated. One method of doing this is to put colon ':' in front of the word. No real word starts with a colon, so the word will not be translated. Later the colon will be removed. All such wordforms that occur in SL and TL should be prefixed with a colon. This can be done with a script located in an appropriate place in the chain of processing.

We will look at the process in phases. In (4) we have a piece of text, which includes sentences in TM.

(4)
Manispaa ya Temeke imepanga kutumia Shilingi 150,000,000 kwa ajili ya kulipa fidia kwa wakazi waliojenga nyumba zao karibu na maeneo ya shule za msingi na ununuzi wa eneo la makaburi. Jina lako nani?
Mkurugenzi wa Manispaa hiyo Bw. Idd Nyundo ameliambia Alasiri kuwa fungu hilo limetengwa kwa ajili ya kuwafidia wakazi waliojenga maeneo yaliyo karibu na shule za msingi za Yombo Vituka na Sabasaba. Unafanya nini?
"Tunataka kuwepo na eneo la wazi kidogo kwa ajili ya viwanja vya michezo katika shule hizo, hivyo baadhi ya nyumba za wakazi hao zitalazimika kubomolewa," akasema Bw. Nyundo. Una watoto wangapi?
Manispaa hiyo itatumia Sh. 29,000,000 kufanya uthamini katika majengo 2231 yaliyopo katika maeneo ya Mbagala, Makangarawe na Yombo Vituka.
Unakwenda wapi? Watu wawili wamekuja. Sina watoto.
Mbali na hilo, Manispaa hiyo itatumia Sh. 10,000,000 katika Mpango wa Miji Endelevu (STP) na Sh. 2,194,000 katika kuweka alama za msingi zitakazorahisisha upimaji wa viwanja. Una umri gani?
Kwa upande wa masoko, manispaa hiyo inatarajia kutumia jumla ya Sh. 23,316,000 katika kurekebisha jengo la katikati ya soko la Temeke Sterio ili kuzuia maji ya mvua yasiingie ndani ya jengo hilo pamoja na kujenga vyoo vya umma katika masoko matatu. Una watoto wangapi?

When we run this text through the TM rules, we get the result as in (5).

(5)

Manispaa ya Temeke imepanga kutumia Shilingi 150,000,000 kwa ajili ya kulipa fidia kwa wakazi waliojenga nyumba zao karibu na maeneo ya shule za msingi na ununuzi wa eneo la makaburi.
What is your name?
Mkurugenzi wa Manispaa hiyo Bw. Idd Nyundo ameliambia Alasiri kuwa fungu hilo limetengwa kwa ajili ya kuwafidia wakazi waliojenga maeneo yaliyo karibu na shule za msingi za Yombo Vituka na Sabasaba.
What :do you :do?
"Tunataka kuwepo na eneo la wazi kidogo kwa ajili ya viwanja vya michezo katika shule hizo, hivyo baadhi ya nyumba za wakazi hao zitalazimika kubomolewa," akasema Bw. Nyundo.
How many children :do you have?
Manispaa hiyo itatumia Sh. 29,000,000 kufanya uthamini katika majengo 2231 yaliyopo katika maeneo ya Mbagala, Makangarawe na Yombo Vituka.
Where :do you go?
Watu wawili wamekuja.
I :do not have children.
Mbali na hilo, Manispaa hiyo itatumia Sh. 10,000,000 katika Mpango wa Miji Endelevu (STP) na Sh. 2,194,000 katika kuweka alama za msingi zitakazorahisisha upimaji wa viwanja.
How old are you?
Kwa upande wa masoko, manispaa hiyo inatarajia kutumia jumla ya Sh. 23,316,000 katika kurekebisha jengo la katikati ya soko la Temeke Sterio ili kuzuia maji ya mvua yasiingie ndani ya jengo hilo pamoja na kujenga vyoo vya umma katika masoko matatu.
How many children :do you have?

The text in (5) is in sentence per line format. We see that seven short sentences have been translated into English. We also see that the verb 'do' has been 'injected' to prevent it from being translated.

Now we pass this text through rule-based translation routines. The result is in (6).

(6)
1. The municipality of Temeke has arranged to use 150,000,000 Shillings because of paying the compensation for the inhabitants who built their houses near the areas of the Primary Schools and purchasing of the area of the graves.
2. What is your name?
3. The director of this Municipality Mr. Idd the Hammer has said to Alasiri that this section has been separated because of compensating the inhabitants who built the areas which are near the Primary Schools of Yombo the Small bushes and Jyly 7th.
4. What do you do?
5. "We want to be there a little clear area because of the plots of the plays in these schools, so some of the houses of these inhabitants will be forced to be destroyed," said Mr. Nyundo.
6. How many children do you have?

7. This municipality will use Sh. 29,000,000 to do valuation in 2231 buildings which are there in the areas of Mbagala, Makangarawe and Yombo the Small bushes.
8. Where do you go?
9. Two people have come.
10. I do not have children.
11. Apart from this, this Municipality will use Sh. 10,000,000 in the Plan of the Sustainable Towns (STP) and Sh. 2,194,000 in placing the basic marks which will make easier assessment of the plots.
12. How old are you?
13. On the side of markets, this municipality hopes to use the total of Sh. 23,316,000 in correcting the building of the middle of the market of Temeke Sterio in order to prevent water of the rain they should not enter in this building together with building the toilets of the public in three markets.
14. How many children do you have?

We see that all sentences have now been translated. It is hard to see which ones were translated with the TM rules and which ones with Salama. For ease of reference, I have numbered the sentences. Translation memory rules have translated sentences no. 2, 4, 6, 8, 10, 12, and 14. They should be correct without editing.

Sentences no. 1, 3, 5, 7, 9, 11, and 13 were translated with the rule-based Salama. We shall take a look at the translation quality of these sentences.

Sentence no. 1:
Translation is grammatically correct.

Sentence no. 3:
The surname of Idd Nyundo was translated as 'the Hammer'. Yhe word needs to be included into the list of proper name candidates. The same concerns the names Vituka and Sabasaba.

Sentence no. 5:
The word cluster 'vivanja vya michezo' is translated by default as 'the plots of the plays'. This should be treated as a multiword expression and translated as 'play grounds'.

Sentence no. 7:
The word order 'Sh. 29,000,000' should be '29,000,000 Sh.'. Again, 'Vituka' is a proper name and should not be translated.

Sentence no. 9:
It is correct.

Sentence no. 11:
There are problems in word order as in 7 above.

Sentence no. 13:

The cluster 'maji ya mvua' is translated with default method as 'water of rain'. This should be treated as a multiword expression with translation 'rain water'.The same concerns 'vyoo vya umma', which should be translated as 'public toilets'.

5 Conclusion

Translation memory and the rule-based Salama translation system can be safely integrated as a single translation system. The accumulation of the translation memory can also been done with a script. This script can be associated to a button on a web interface to make the translation process convenient. Whe translation memory and the rule-based translation system are kept separate from each other, tracing mistakes in the system becomes straightforward.