Steps Toward Accurate Machine Translation

Kanlaya Naruedomkul, Nick Cercone

Department of Computer Science
University of Regina
Regina, Saskatchewan, Canada
e-mail: {kanlaya, nick}@cs.uregina.ca

Abstract. A novel multi-phase architecture for an accurate machine translation system is proposed. The system is divided into three phases: ***quick and dirty machine translation*** (QDMT), ***conceptual comparison*** and ***repair and iterate***. QDMT generates the appropriate translation candidates (TCs) in the target language for the input sentence in the source language. Next, the system compares the meaning of the TC with that of the input sentence. If there is dissimilarity in meaning between the TC and the input sentence, the most appropriate TC will then be "repaired". To demonstrate this approach, a translation system which translates from English to Thai has been developed. In this paper, QDMT is described and initial experiments with QDMT are presented. Some concluding remarks are made with respect to completion of the first phase.

1. Introduction

We propose a novel architecture of *Generate and Repair Machine Translation* (GRMT, Figure 1). This system is designed for accurate translations primarily, speed secondarily. The first phase, quick and dirty machine translation (QDMT) has been fully conceptualized and is partially implemented. The result of this initial phase is to generate the most appropriate translation candidate (TC) into the target language (TL) of the source language (SL) utterance which subsequently can easily be developed into a "correct, accurate" translation. In the second phase, conceptual comparison, we will employ sophisticated head driven phrase structure grammar (HPSG) [Pollard and Sag, 1987] parsers for the source and target languages in order to conceptually compare the parser's outputs. When the conceptual comparison phase indicates a discrepancy in meanings beyond an acceptable threshold, we then "repair" and "iterate" the most likely translation using a variety of techniques, many of which remain to be articulated, until we are confident that the translation is accurate.

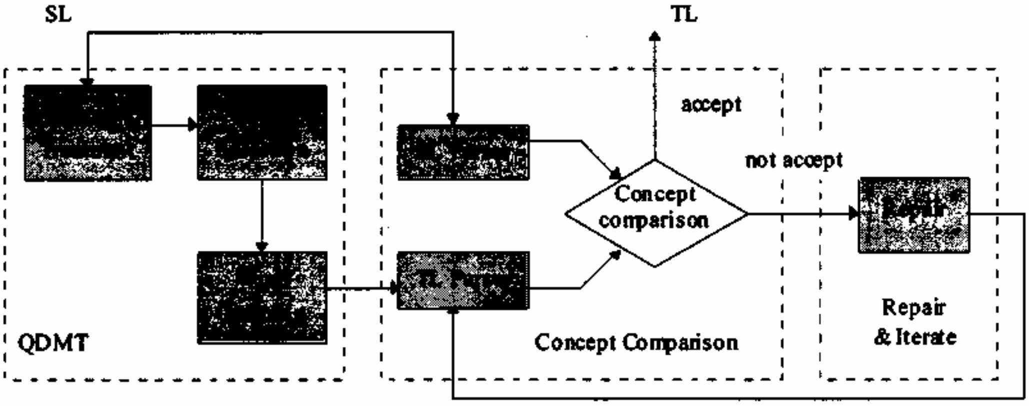

Figure 1. GRMT Architecture

Three widely articulated strategies for developing machine translation systems include Direct

MT, a "word-to-word" replacement approach whose accuracy depended primarily on the bilingual dictionary, Transfer MT, which "transferred" the information of SL to TL by analyzing and representing the SL in an internal form and then generating the TL (the accuracy of which depended primarily upon the sophistication of the transfer module), and Interlingual MT, which used an intermediate language-independent conceptual representation between SL and TL called Interlingual which permitted independent parsing and generation of SL and TL. Table 1 illustrates some comparisons between these three approaches. In addition, nonlinguistic information strategies have become popular, e.g., knowledge-based strategies [Nirenburg et al.,1992], statistical strategies [Brown et al., 1992], example-based MT [Jones, 1996], etc.

Attributes	Direct MT	Transfer MT	Interlingual MT
Factors of Accuracy	dictionary and mapping rule	transfer rule	concept representation
Levels of Linguistic Analysis	word	meaning	concept
Intermediate Representation	no Representation	language dependent representation	language independent represenation
Modularlity	depends on system design	depends on system design	analysis and generation modules independent
Multilingual System (add the nth language to the (n-1) languages system)	needs 2(n-1) mapping rules	needs (n-1) analysis, (n-1) generation and 2(n-1) transfer	needs 1 analysis and 1 generation
Extendibility (in terms of integrating new language to system)	needs mapping rules	needs analysis, generation and tran sfer	needs analysis and/or generation

Table 1. Comparison of the Three Approaches

There are advantages and disadvantages of each approach. The efficiency inherent in the direct strategy is limited because linguistic elements of the languages are considered only at the morphology level. The transfer of structure information in the transfer strategy may cause inaccuracy because of the different structures between languages and thus may lose some information during processing. Table 2 illustrates this problem of losing information. The second column in French represents the translation of the first column from English[1]. Each translation fails to maintain close fidelity to the meaning of the original sentence. The translations are incorrect in both words selected and in grammar. The third column shows the translation back into English of the French translation in the second column[2]. The third and the first columns should be the same. In addition, the transfer approch does not appear appropriate for multilingual systems. Of the three approaches, although the interlingual approach appears to be the most attractive, the interlingual representation is still an ideal. To overcome shortcomings of the direct and transfer strategies, and to increase the accuracy of the overall translation system while avoiding the difficulty in establishing an interlingual representation, GRMT is proposed. By performing a *conceptual comparison* phase, GRMT ensures that the final generated translation retains the meaning of the original sentence.In this paper, QDMT is described and outlined by using English as the SL, and Thai as the TL. The latter phases will be discussed in a subsequent paper.

2. QDMT Architecture

QDMT was designed around two simple notions: first, the more accurate QDMT generates a TC, the less works is required in the latter phases; and second, generating the TC must be done quickly. Therefore, QDMT generates a TC by considering the difference between language

[1, 2] These translations were provided by a commercial MT system.

English	Translation to French	Translation back to English
1. Hatchery officials are having to teach the fish to like worms.	1. *Les fonctionnaires de l'incubateur doivent apprendre le poisson pour aimer des vers.*	1. Civil servants of the incubator must learn fish for aimerdes toward.
2. It does not matter if you are born in a duck yard.	2. *Il n'importe pas si vous naissez dans un jardin du canard.*	2. He/it doesn't import if you are born in a garden of the duck.
3. Only a life lived for others is a life worth while.	3. *Seulement une vie pour les autres vaut pendant que.*	3. Only a life for other is worth while.
4. I never think of the future.	4. *Je ne pense jamais du futur.*	4. I never think the future.
5. You can take a fish to school, but you can not make them think.	5. *Vous pouvez prendre un poisson pour scolariser, mais vous ne pouvez pas les faire penser.*	5. You can take a fish to school, but you don't can pasles to make think.
6. It's no go.	6. *Il n'est pas aucun entrain.*	6. Him estpas no liveliness.
7. Never mind.	7. *Ne faites jamais attention.*	7. Don't make attention never
8. BOISE, Idaho (AP)- Trout in Idaho are not just swimming in schools- they are going to school.	8. *BOISE, Idaho (AP) - Truite dans Idaho ne nage pas dans les écoles juste - theyare aller scolariser.*	8. BOISE, Idaho (AP) - Trout in Idaho doesn't swim rightly in schools- theyare to be going to school.

Table 2. Translation Examples by a Commercial System

pairs in terms of syntax and semantics without performing any sophisticated analysis. QDMT performs its task by judiciously selecting a few efficient heuristics, constraints, and semantic principles to apply when appropriate. QDMT first considers TL words which correspond to all possible meanings of each SL word. The most appropriate TL word is selected by applying a semantic relationship between words and then the selected words are rearranged according to the grammar of the TL. QDMT comprises three modules as shown in Figure 2, *word treatment*, *word selection*, and *word ordering*.

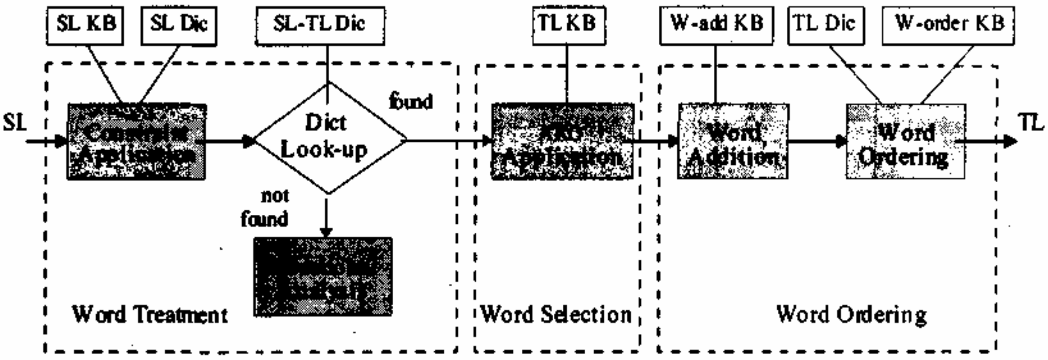

Figure 2. QDMT Architecture

2.1 Word Treatment

There are two steps performed by this module: SL Constraint Application and Dictionary look-up. Word Treatment requires Inflectional rules, the SL dictionary and the SL-TL Dictionary.

2.1.1 SL Constraint Application

To narrow the scope of possible TL words which correspond to each SL word, some characteristics of the SL which differ from those of the TL are incorporated in this step as constraints. In this study, the following characteristics of English are considered.

1. Auxiliary Verb Constraint

Auxiliary verbs in English are needed in many cases, e.g., in front of adjectives or negative "not", but they are not used in Thai for the same expression as shown in following examples:

English		Thai
be + adjective	→	adjective
I am glad	→	ฉัน ดีใจ (chân- I) (diicaj- glad)[3]
do + not	→	not
He does not eat	→	เขา ไม่ กิน (khaw- he) (mâj - not) (kin- eat)
There + be	→	to have
There is a book.	→	มี หนังสือ หนึ่ง เล่ม (mii -'There-is') (naŋsy̌y -book) (ny`ŋ -a) (lêm- unit)

2. Present Continuous and Present Perfect (Continuous) Form Constraint
   In English, the inflection "ing" form of a verb is needed after an auxiliary verb to describe an action that is going on at the moment of speaking. However, the Thai language does not have this inflection, therefore, the word "กำลัง -kamlaŋ" is used to describe the same action without changing the verb form, for example:

be + V ing	→	กำลัง + V
I am swimming.	→	ฉัน กำลัง ว่ายน้ำ (chân- I) (kamlaŋ- ing) ('wâaj-náam'- swim)

(Exception: interesting, ...)

3. Passive Voice Constraint
   Another constraint is the treatment of passive voice, which is used in English but in Thai, passive voice is commonly used to convey an unpleasant situation. For example:

He was killed.	→	เขา ถูก ฆ่า (khaw - he) (thùug - passive) (khâa - kill)
He was arrested.	→	เขา ถูก จับ (khaw - he) (thùug - passive) (càb - arrest)
The book was taken by him.	→	หนังสือ ถูก เขา เอา ไป (naŋsyy -book) (thùug -passive) (khaw -him) (?aw -take) (paj - modifying) เขา เอา หนังสือ ไป (khaw -him) (?aw -take) (naŋsy̌y -book) (paj - modifying)

The word "ถูก" denotes the passive voice in Thai. In the third example, the first Thai sentence, which is passive, is grammatically correct but it is not the way Thai people convey this expression. They will use an active voice rather than passive voice in this situation as shown in the second sentence.

2.1.2 Dictionary Look-up

Each word of the input string which is an output of the previous step will be used as a keyword to search for the corresponding word in the TL. If the keyword used can be found in the bilingual dictionary, all possible corresponding TL words will be attached to that SL word. If the keyword cannot be found, inflectional analysis is performed before re-searching. Inflectional analysis provides information about tense, plural, present participle and comparison for such input words. It also indicates the part of speech of the word, for example, verb for tense and present participle, noun for plural and adjective for comparison. This information is useful in the word selection step and also reduces the size of dictionary required and the search time.

[3] Phonetic transcription of Thai word and gloss.

In this study, English is regarded as SL and its following inflectional forms will be considered: Past and Perfect Tenses forms (-ed), Plural form (-s, -es), Present participle (-ing), Comparative and Superlative forms (-er, -est). An output of this step is a string of SL words which are in root form plus the information of the original form, for example:

took	→	take	(past,v)	standing	→	stand	(ing,v)
eggs	→	egg	(plural,n)	hardest	→	hard	(sup,adj)

2.2 Word Selection

To select the most appropriate TL word for each SL word, a semantic relationship between words is considered. We have designed the semantic relationship by using the "A Kind of" (AKO) [Tantisawetrat et al., 1991] slot from the CICC Multilingual Machine Translation project [CICC, 1995]. This relation indicates which word can occur with which word in the same sentence. For example, in "Five days", the word "day" can be translated as "กลางวัน - klaaŋwan" or "วัน -wan" in Thai "กลางวัน" means "the time between sunrise and sunset, AKO number 2-7-2-2" while "วัน" means "a period of 24 hours, AKO number 2-7-2-2" and it also can be used as a classifier in Thai which has AKO "2-10-1". The word "five" is translated to "ห้า -hâa" with AKO "1-1-2-1-2-1" and the semantic relationship shows that this word can occur with the word which has AKO "2-10-1". So, the appropriate word "วัน" is selected as indicated by the AKO value of "five".

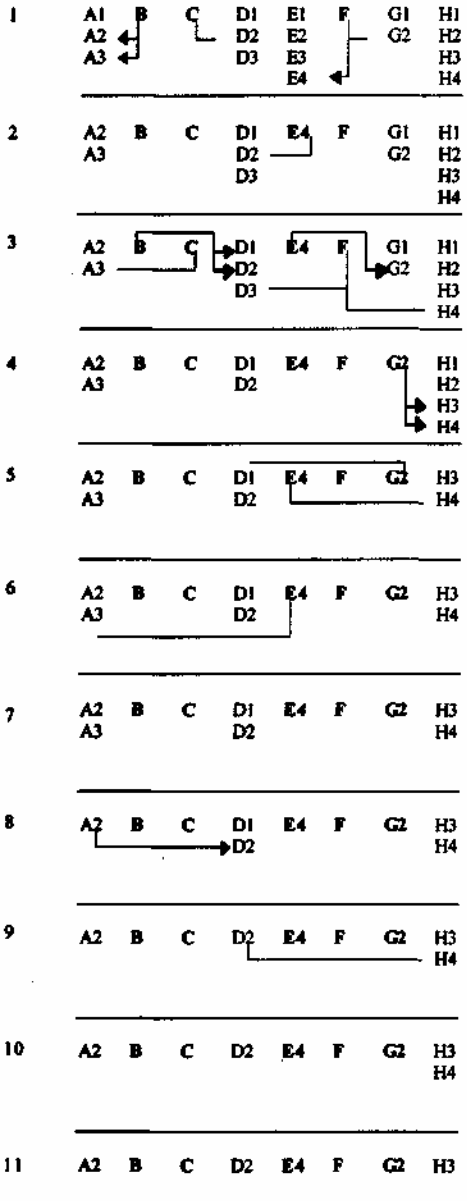
1 A1 A2 A3 B C D1 D2 D3 E1 E2 E3 E4 F G1 G2 H1 H2 H3 H4
2 A2 A3 B C D1 D2 D3 E4 F G1 G2 H1 H2 H3 H4
3 A2 A3 B C D1 D2 D3 E4 F G1 G2 H1 H2 H3 H4
4 A2 A3 B C D1 D2 E4 F G2 H1 H2 H3 H4
5 A2 A3 B C D1 D2 E4 F G2 H3 H4
6 A2 A3 B C D1 D2 E4 F G2 H3 H4
7 A2 A3 B C D1 D2 E4 F G2 H3 H4
8 A2 B C D1 D2 E4 F G2 H3 H4
9 A2 B C D2 E4 F G2 H3 H4
10 A2 B C D2 E4 F G2 H3 H4
11 A2 B C D2 E4 F G2 H3

Figure 3. Word Selection Step

As well as applying the semantic relationship between words, the selection step continues. Based on connections between words: words in close proximity are considered to have stronger connections. The selection steps are given as follows, where $1 \leq i, j \leq n$, $2 \leq m \leq n\text{-}1$, n is the number of words in the SL sentence:

1. If each x_i and x_j $(i \neq j)$ has a unique meaning
- then use x_i to determine the meaning of x_{i-1} and x_{i+1} if each x_{i-1}, x_{i+1} has more than one meaning.
- and use x_j to determine the meaning of x_{j-1} and x_{j+1} if each x_{j-1}, x_{j+1} has more than one meaning.
- After these determinations, if x_{i+1} and x_{j-1} (where $i+1 = j-1$) have different meanings, keep all of them.

Repeat this process, adjusting i and j appropriately, until all (new) unique meaning words have been treated.

2. If each x_i and x_j $(i \neq j)$ has a unique meaning

- then use x_i to determine the meaning of x_{i-m} and x_{i+m} if each x_{i-m}, x_{i+m} has more than one meaning.
- and use x_j to determine the meaning of x_{j-m} and x_{j+m} if each x_{j-m}, x_{j+m} has more than one meaning.
- After these determinations, if x_{i+m} and x_{j-m} (where $i+m = j-m$) have different meanings, keep all of them.

If new unique meaning words occur in this step, repeat step 1 for these new unique meaning words only.

Repeat this process, adjusting i and j appropriately,until all (new) unique meaning words have been treated.

Step 2 begins with $m = 2$, and we repeat this step with $m = 3, 4, 5, \ldots$ until every word has been assigned a unique meaning or this process cannot go any further. If after this process has been exhausted there are still some words which have more than one meaning e.g., x_k and x_l, $(l > k)$, consider x_k. Select the first dictionary meaning of x_k and repeat step 1 for this new unique meaning word.

Figure 3 illustrates an example of the selection step process. In this illustration an eight word sentence is examined with words A, B, ..., H. Each of B, C and F has a unique in this example.

First consider A and C which are on the left and right of B. A has three meanings, A1, A2 and A3, so we select the meaning of A by considering the semantic relation between A and B. Suppose A2 and A3 have meaningful relations with B, then keep both of them. C already has only one meaning so skip it. Next, we consider B and D to resolve against C in a similar manner.

It is shown in Figure 3 that the meaning of E can be selected (constrainted) by F. Then we consider the words that are adjacent to E. Once we finish considering a word which is next to the word which has one meaning, we will consider the word which is two words apart from it and repeat the same process; we keep repeating this process until every word has only one unique meaning or we cannot go any further. Whenever we are left with words without a unique meaning (Figure 3, part 7), we consider the first word which still has more than one meaning and then select the first dictionary meaning for it and use that meaning to constrain the meanings of the rest, e.g., A2 in Figure 3 part 7. It is assumed that these meanings are ordered by the frequency of usages in the dictionary.

2.3 Word Ordering

There are syntactic level differences between English and Thai, even though the typical sentence structure of both languages is basically the same. The typical sentence contains subject, verb and object in that order, e.g.,

He^1 buys^2 a book^3. เขา^1 ซื้อ^2 หนังสือ^3
(khaw- he) (sy'y - buy) (naŋsy'y - book)

However, a type of sentence which contains no subject is also used, e.g.,

There is^1 a book^2 on^3 the desk^4. มี^{1} หนังสือ^2 บน^3 โต๊ะ^4
(mii - There-is) (naŋsy'y - book) (bon - on) (tó? - desk)

Also the head usually comes before the attribute. For example:

My^1 father^2	พ่อ^2 ของผม^1 (phɔ̂ɔ -father) (khɔ̌ɔŋ-phom -my)
The^1 red^2 book^3.	หนังสือ^3 สีแดง^2 เล่ม นั้น^1 (naŋsy̌y - book) (sǐidɛɛŋ - red) (lêm - unit) (nán - the)

Modality is used as an auxiliary verb or adverb, e.g.,

He^1 will^2 not^3 go^4 home^5.	เขา^1 จะ^2 ไม่^{3} ไป^4 บ้าน^5 (khaw- he) (càʔ - will) (mâj - not) (paj - go) (bâan - home)

A sentence may contain a series of verbs, only one of these verbs is a head word and the rest are *modifications or prepositions.*

He^1 walks^2 to^3 school^4.	เขา^1 เดิน^2 ไป^3 โรงเรียน^4 (khaw- he) (dəən - walk) (paj - go) (rooŋrian - school)

Any selected word which contains information about tense, plurality, present participle and comparison will be treated in this step as well. The Thai language does not have an inflection resulting from verb agreement, number or tense as in the English language. It does not matter what the subject is, the word "buy" is translated as "ซื้อ - syy". Even in the past tense, the word "bought" is also translated as "ซื้อ". However, tense can be shown by a modifying verb or by indicating the time frame, while number can be expressed by using a classifier, if necessary. For example, the word "หลาย - laaj" indicates that there are more than one book and "เล่ม - lêm" is a classifier.

I buy a book.	ฉัน ซื้อ หนังสือ (chân- I) (sy′y - buy) (naŋsy̌y - book)
He buys a book.	เขา ซื้อ หนังสือ (khaw- he) (sy′y - buy) (naŋsy̌y - book)
He buys books.	เขา ซื้อ หนังสือ (หลาย เล่ม) (khaw- he) (sy′y - buy) (naŋsy̌y -book) (la`aj -plural) (lêm- unit)
He bought a book.	เขา ซื้อ หนังสือ แล้ว (khaw- he) (sy′y - buy) (naŋsy̌y - book) (lɛ′ɛw - past tense)

Classifiers indicate the unit of a countable noun. Classifiers play an important role in noun constructions which express a quantity or modify a noun. The types of classifier are not restricted to any kind of expression [Sornlertlamvanich, 1994]. For example:

two books	หนังสือ สอง เล่ม (naŋsy̌y - book) (sɔ̌ɔj - two) (lêm- classifier)
a cat	แมว ตัว หนึ่ง (mɛɛw - cat) (tua - classifier) (ny`ŋ - a)
one cat	แมว หนึ่ง ตัว (mɛɛw - cat) (ny`ŋ - one) (tua - classifier)
some animals	สัตว์ บาง ชนิด (sàd - animal) (baaŋ - some) (chaníd - classifier)

In many systems, the ordering step has been performed by generating valid combinations of words and analyzing (parsing) to see whether the combination is grammatically correct. If it is, that combination will be selected as an output; if not, another combination is tried until the grammatically correct sentence is found. This algorithm is time consuming and requires significant effort. In this project, the selected words are ordered according the rules without performing any analysis. These ordering rules are generated from a number of TL examples consistent with the Thai verb pattern (Table 3). The ordering is considered based on the *subcategory* information of the word to more narrowly refine the functions of the word. Examples of ordering rules are shown in Figure 4. Names in the ordering relations and Thai verb patterns are given in the legend.

Verb Pattern	Samples
SUB + V	เขา เดิน he walk
V + DOB	มี แผ่นดินไหว exist earthquake
SUB + V + ADV	เขา รบ กัน he fight each other
SUB + V + AUX	เขา ลุก ขึ้น he wake up
SUB + V + PP	เขา แต่งงาน กับ เธอ he marry with her
SUB + V + DOB + PP	เขา ซื้อ ขนม ให้ ลูก he buy sweet to child
SUB + V + DOB	เขา พิมพ์ รายงาน he type report
SUB + V + DOB + IOB	เขา ให้ เงิน เธอ he give money her
SUB + V + COMP	เขา โกรธ ที่ เรา มา สาย he angry that we come late
SUB + V + DOB + COMP	เขา บอก ฉัน ว่า เขา ง่วง he tell me that he sleepy
SUB + V + PP + COMP	เขา บอก กับ เธอ ว่า เขา ง่วง he tell with her that he sleepy

Table 3. Thai Verb Pattern[4]

Legend:

naln	- adverb like noun
vact	- active verb
neg	- negation
ncmn	- cardinal number
pprs	- personal pronoun
ccrg	-coordinating conjuction
nlbl	- label noun
ddac	- determiner
cnit	- unit classifier
nclt	- collective noun
rpre	- preposition
csbr	- subordinate conjunction
vatt	- attribute verb
vsta	- stative verb
xvam	- auxiliary verb
xvbm	- auxiliary verb
SUB	- Subject of sentence
DOB	- Direct object
COMP	- Complement
V	- Verb
PP	- Prepositional phrase
ADV	- Adverb
AUX	- Auxiliary verb
IOB	- Indirect obiect

```
order_relation(naln,[ncmn]).
order_relation(ncmn,[ddac,cnit,vatt,ncmn,nlbl,ccrg,xvam,pprs]).
order_relation(pprs,[vsta,xvam,neg,vact,xvbm]).
order_relation(nclt,[ccrg]).
order_relation(cnit,[rpre]).
order_relation(vact,[ncmn,rpre,csbr,pprs,nclt]).
order_relation(neg,[xvam,vsta,vact]).
```

Figure 4. Examples of Ordering Rules

3. Dictionaries

There are three types of dictionary used in our approach, the SL dictionary, TL dictionary and a bilingual dictionary. Entries in the SL and TL dictionaries can be single word, some inflected and derived forms which cannot be easily handled by rules. Compound words are also included. Each entry contains morphological, syntactic and semantic information. Examples of entries in the three Dictionaries are shown below in Figure 5. The Thai dictionary entry contains word form and word subcategory. The English dictionary contains the category is used in the inflectional analysis step. The Bilingual dictionary contains the English entry and all corresponding Thai words and an AKO number for each Thai word, e.g., the word "dream" in English has three Thai words which express differences in meaning and usage. All Thai words which correspond to each English entry are ordered based on the frequency of usage (in real life). The first meaning will be selected once the constraint and AKO fail.

[4] Verb Pattern indicates the syntactic structure of verb in the Thai language. (CICC, Technical Report:Thai Generation System).

```
        entry(day,[ako('วัน','2-10-1', '2-7-2-2'),ako('กลางวัน','2-7-2-2')]).
        entry(dream,[ako('ฝัน','2-2'),ako('ความฝัน','2-6-1'),ako('การฝัน','2-3-9')]).
        entry(duckling,[ako('ลูกเป็ด','1-1-2')]).
        entry(five,[ako('ห้า','1-1-2-1-2-1')]).
        entry(glad,[ako('ดีใจ','2-2-7'),ako('ความยินดี','2-6-1')]).                  SL-TL Dic

entry_e(day,[n,clas]).                              entry_t('วัน',[cnit]).
entry_e(dream,[v,n]).                               entry_t('กลางวัน',[naln]).
entry_e(duckling,[n]).                              entry_t('ฝัน',[vsta]).
entry_e(five,[n]).                                  entry_t('ความฝัน',[ncmn]).
entry_e(glad,[adj]).            SL Dic              entry_t('การฝัน',[nnom]).           TL Dic
```

Figure 5. Examples of Dictionaries

4. Examples of QDMT

In this section we present the results of applying QDMT to some example sentences. The input symbols of Example 1 are shown in the first column labelled input (Table 4). The second column illustrates the output after applying the constraints e.g., "are used" triggers the "passive voice" constraint, "do not" triggers the "negative" constraint and "are talking" triggers "present continuous" constraint. Each word in this column is used as a keyword to search for the corresponding words in Thai. The word "symbols" is analyzed in terms of "plurality" before it can be found in a bilingual dictionary as described in section 2.1. All possible meanings of each input word are shown in the third column. Some of them have more than one meaning e.g., "symbol", "when", "know", etc. The appropriate meaning of "use", "know" and "about" can be selected by considering the semantic relationship between words and the choice for each of these is shown in the fourth column. However, the appropriate words for "symbol" and "when" cannot be selected in the same manner because all possible meanings of each word have the same AKO number. Therefore, the first meaning which appears in the bilingual dictionary of each is selected. All selected words are shown in the last column. Before performing the ordering step, the word "ว่า - wâa" is added to combine clauses. The word "ว่า" is a translation of the word "that" which is omitted in this sentence, however, it cannot be omitted in Thai otherwise the translation in Thai will be grammatically incorrect. In Example 1, the generated TC formed the "Correct translation" for the input sentence without the necessity of performing any correction.

Input	Constraint Application	Dic Lookup& Inflect Analysis	Word Selection	Selected Word
Algebraic	Algebraic	ทางพีชคณิต	ทางพีชคณิต	ทางพีชคณิต
symbols	symbols	สัญลักษณ์ เครื่องหมาย	สัญลักษณ์ เครื่องหมาย	สัญลักษณ์
are	passive	ถูก	ถูก	ถูก
used	use	ใช้ ประโยชน์	ใช้	ใช้
when	when	เมื่อ, เมื่อไหร่,ขณะที่	เมื่อ เมื่อไหร่ ขณะที่	เมื่อ
you	you	คุณ	คุณ	คุณ
do	do	-	-	-
not	not	ไม่	ไม่	ไม่
know	know	รู้ รู้จัก	รู้	รู้
what	what	อะไร	อะไร	อะไร
you	you	คุณ	คุณ	คุณ
are	ing	กำลัง	กำลัง	กำลัง
talking	talk	พูด	พูด	พูด
about	about	ประมาณ, เกี่ยวกับ, รอบๆ	เกี่ยวกับ	เกี่ยวกับ

Table 4 QDMT Steps Applied to the Sentence of Example 1

Example 1: Algebraic symbols are used when you do not know what you are talking about.

TC: สัญญลักษณ์ ทางพีชคณิต ถูก ใช้ เมื่อ คุณ ไม่ รู้ ว่า คุณ กำลัง พูด เกี่ยวกับ อะไร

CT: สัญญลักษณ์ ทางพีชคณิต ถูก ใช้ เมื่อ คุณ ไม่ รู้ ว่า คุณ กำลัง พูด เกี่ยวกับ อะไร
(sanjalág -symbol) (thaaŋ-phichakhanίd -algebraic) (thùug -passive) (chái -use) (my^a -when) (khun -you) (mâj -not) (rúu -know) (wâa -connective) (khun -you) (kamlaŋ -ing) (phûud -talk) ('kiàw-kàb' -about) (araj -what)

In generating the TC for Example 2, once all words are selected (Table 5), the words "ที่. . . จะ (thîi...cà?)" must be added to clarify tense. These additions are necessary because the preposition "before" shows that the "release" action will be taken in future. The word "ทั้งหลาย (tháŋ-laˇaj)" is also added to show the "plurality" of "trout". This TC is exactly the same as the correct translation. QDMT generates the appropriate TC not only for sentences but also for phrases as it is shown in example 2.

Input	Constraint Application	Dic. Lookup& Inflect Analysis	Word selection	Selected Word
Five	Five	ห้า	ห้า	ห้า
days	days	day+S วัน กลางวัน	วัน	วัน
before	before	ก่อน	ก่อน	ก่อน
the	the	นั้น	นั้น	นั้น
trout	trout	ปลาเทราท์	ปลาเทราท์	ปลาเทราท์
are	passive	ถูก	ถูก	ถูก
released	release	ปล่อย	ปล่อย	ปล่อย

Table 5 QDMT steps applied to the phrase of Example 2

Input	Constraint Application	Dic Lookup& Inflect Analysis	Word Selection	Selected Word
You	You	คุณ	คุณ	คุณ
can	can	สามารถ, กระป๋อง, ทำเป็นกระป๋อง, บรรจุกระป๋อง, คุก, หองน้ำ	สามารถ	สามารถ
take	take	เอา, หยิบ, จับ, ลาก, พา	เอา, หยิบ, จับ, ลาก, พา	เอา
a	a	ตัวหนึ่ง, อันหนึ่ง, คนหนึ่ง	ตัวหนึ่ง	ตัวหนึ่ง
fish	fish	ปลา	ปลา	ปลา
to	to	ไป, ไปถึง, ก่อน, จนกระทั่ง, ด้วยกับ, เป็นส่วนของ	ไป, ไปถึง, ด้วยกับ, เป็นส่วนของ	ไป
school	school	โรงเรียน, ฝูง, สอน, อบรม, ซ้อม	ฝูง	ฝูง
but	but	แต่, นอกจาก	แต่	แต่
you	you	คุณ	คุณ	คุณ
can	can	สามารถ, กระป๋อง, ทำเป็นกระป๋อง, บรรจุกระป๋อง, คุก, หองน้ำ	สามารถ	สามารถ
not	not	ไม่	ไม่	ไม่
make	make	ทำให้	ทำให้	ทำให้
them	them	พวกเขา	พวกเขา	พวกเขา
think	think	คิด, เห็นว่า, นึก	คิด, เห็นว่า, นึก	คิด

Table 6 QDMT Steps Applied to the Sentence of Example 3

Example 2: Five days before the trout are released.

TC: ห้า วัน ก่อน ที่ ปลาเทราท์ ทั้งหลาย นั้น จะ ถูก ปล่อย

CT: ห้า วัน ก่อน ที่ ปลาเทราท์ ทั้งหลาย นั้น จะ ถูก ปล่อย
(hâa-five) (wan-day) (kə`ən-before) (thîi -modifying) ('plaa- thráaw'-trout) ('tháŋ-laˇaj'- plurality) (nán-the) (cà?-modifying) (thùug-passive) (plə`əj -release)

In the TC of example 3 (Table 6), the translation of the word "school" is "ฝูงปลา - fuuŋ-plaa" which means "a large group of one kind of fish or certain other sea animals swimming together" [Longman, 1992]. This selection was based on the semantic relationship between words "fish" and "school" which is correct in linguistic meaning. However, the translation of "school" in this expression should be "โรงเรียน -rooŋrian" which means "a place of education for children". [Longman, 1992] because of the speaker intention. It was only one word which was translated inappropriately, and it will be corrected in the *Repair and Iteration* phase.

Example 3: You can take a fish to school, but you can not make them think. (sic, as appearing in the 26 July 1996 Regina Leader-Post newspaper)

TC: คุณ สามารถ เอา ปลา ตัวหนึ่ง ไป <u>ฝูงปลา</u>, แต่ คุณ ไม่ สามารถ ทำให้ พวกเขา คิด

(khun -you) (saamâad -can) (?aw -take) (plaa -fish) ('tua-ny`ŋ' -a) (paj -to) ('<u>fuuŋ-plaa' -school</u>) (tɛ`ɛ - but) (khun -you) (mâj -not) (saamâad -can) (tham-hâj -make) ('phûag-khaw' -kìd)

CT: คุณ สามารถ เอา ปลา ตัวหนึ่ง ไป <u>โรงเรียน</u>, แต่ คุณ ไม่ สามารถ ทำให้ พวกเขา คิด

(khun -you) (saamâad -can) (?aw -take) (plaa -fish) ('tua-ny`ŋ' -a) (paj -to) (<u>rooŋrian-school</u>) (tɛ`ɛ - but) (khun -you) (mâj -not) (saamâad -can) (tham-hâj -make) (phûag-khaw -kìd)

Example 4: When I was an ugly duckling he thought I never dreamed I could be so happy.

TC: เมื่อ ฉัน เป็น ลูกเป็ด ตัวหนึ่ง เขา คิด ฉัน ไม่เคย ฝัน ว่า ฉัน สามารถ <u>เป็น</u> ความสุข ขี้เหร่ มาก

(my^a -when) (chân -I) (pen -was) ('lûug-pèd' -duckling) ('tua-nyŋ' -an) (khaw -he) (kìd -thought) (chân -I) ('mâj-khəəj' -never) (faˇn -dream) (wâa -connective) (chân -I) (saamaâd -could) (<u>pen -be</u>) ('khwaam-sùg' -happy) (*khîirèe -ugly*) (mâag -so)

CT: เมื่อ ฉัน เป็น ลูกเป็ด ขี้เหร่ ตัวหนึ่ง เขา คิด ฉัน ไม่เคย ฝัน ว่า ฉัน สามารถ <u>มี</u> ความสุข มาก

(my^a - when) (chân -I) (pen -was) ('lûug-pèd' -duckling) (*khîirèe -ugly*) ('tua-nyŋ' -an) (khaw -he) (kìd -thought) (chân -I) ('mâj-khəəj' -never) (faˇn -dream) (wâa -connective) (chân -I) (saamâad - could) (<u>mii -be</u>) ('khwaam-sùg' -happy) (mâag -so)

Input	Constraint Application	Dic Lookup& Inflec Analysis	Word Selection	Selected Word
When	When	เมื่อ, เมื่อไหร่, ขณะที่	เมื่อ, เมื่อไหร่, ขณะที่	เมื่อ
I	I	ฉัน	ฉัน	ฉัน
was	was	be ใช่, เป็น, อยู่, คือ, เป็นอยู่, มี	เป็น	เป็น
an	an	ตัวหนึ่ง, อันหนึ่ง, คนหนึ่ง	ตัวหนึ่ง	ตัวหนึ่ง
ugly	ugly	ขี้เหร่, น่าเกลียด	ขี้เหร่	ขี้เหร่
duckling	duckling	ลูกเป็ด	ลูกเป็ด	ลูกเป็ด
he	he	เขา	เขา	เขา
thought	thought	คิด, เห็นว่า, นึก	คิด, นึก	คิด
I	I	ฉัน	ฉัน	ฉัน
never	never	ไม่เคย	ไม่เคย	ไม่เคย
dreamed	dreamed	ฝัน, การฝัน, ความฝัน	ฝัน	ฝัน
I	I	ฉัน	ฉัน	ฉัน
could	could	สามารถ	สามารถ	สามารถ
be	be	ใช่, เป็น, อยู่, คือ, เป็นอยู่, มี.	เป็น	เป็น
so	so	มาก	มาก	มาก
happy	happy	ความสุข	ความสุข	ความสุข

Table 7 QDMT Steps Applied to the Sentence of Example 4

Again, the word "ว่า -wâa" is added before the ordering step can be performed to combine clauses. The word "be" should be translated as "มี -mii" in Example 4 (Table 7). But it is translated as "เป็น -pen" which is inaccurate because AKO numbers indicate that all these possible meanings can occur with all words which are in close proximity. So, QDMT selects the first meaning which was found in the *Dictionary Look-up* step. Another mistake in this translation is the order of the word "ขี้เหร่ -khîrèe". This ordering is wrong because the ordering is not yet complete.

We have not tested QDMT exhaustively as yet. The four examples we present are illustrative of the performance of this current QDMT prototype and were chosen to illustrate the points we discussed. Further progress is expected in this QDMT phase, some of which is discussed in the next section.

5. QDMT is the First Phase But ...

We illustrate another aspect of QDMT's performance in contrast with a commercial MT system on selected sentences. The example sentences were selected before QDMT was built and they were run against the commercial system to take advantage of a "free-usage" offer. We did it so in order to gauge the accuracy of modern commercial systems.

Subsequently, the QDMT prototype was constructed and we remembered our earlier "experiment" of the translation by commercial MT system. We then attempted to translate these same sentences using QDMT, Table 8 depicts the results.

In Table 8, the second column illustrates the translation in French of the first column from English. The translation which was provided by the commercial MT system is shown numbered and the correct translation is shown directly below in italic font. Each generated translation is incorrect in both words selected and grammar.

English	Translation in French	Translation in Thai
1. The wheat was yellow.	1. Le blé érait jaune. *Le blé était jaune.*	1. ข้าวสาลี นั้น สีเหลือง *ข้าวสาลี นั้น สีเหลือง*
2. Here the stork marched about on his long red legs.	2. Ici la cigogne a marché environ sur ses longues jambes rouges, en discuter Egyptian. *Ici, la cigogne se mit à tourner en rond sur ses longues pattes rouges.*	2. ที่นี่ นกกระสา นั้น เดินแถว รอบๆ ที่ ขา ยาว ของเขา สีแดง *ที่นี่ นกกระสา นั้น เดินแถว รอบๆ ด้วย ขา ยาว สีแดง ของเขา*
3. When I was an ugly duckling he thought I never dreamed I could be so happy.	3. Quand ètais un caneton laid, il pensait, je n'ai jamais rêvé que je pourrais ètre si heureux. *Jamais je n' aurais* rêvé *, lorsque j' étais un rilain petit canard, que je pourrais être si heureux, pensa-t-il.*	3. เมื่อ ฉัน เป็น ลูกเป็ด ตัวหนึ่ง เขา คิด ฉัน ไม่เคย ฝัน ฉัน สามารถ เป็น ความสุข ขี้เหร่ มาก *เมื่อ ฉัน เป็น ลูกเป็ด ขี้เหร่ ตัวหนึ่ง เขา คิด ฉัน ไม่เคย ฝัน ฉัน สามารถ มี ความสุข มาก*
4. You can take a fish to school but you can not make them think.	4. Vous pouvez prendre un poisson pour scolariser, mais vous ne pouvez pas les faire penser. *Vous pouvez emmener un poisson à l' école cependent il est impossible gue vous l'obligez de penser.*	4. คุณ สามารถ เอา ปลา ตัวหนึ่ง ไป ฝูง แต่ คุณ ไม่ สามารถ ทำให้ พวกเขา คิด *คุณ สามารถ เอา ปลา ตัวหนึ่ง ไป โรงเรียน แต่ คุณ ไม่ สามารถ ทำให้ พวกเขา คิด*
5. Five days before the trout are released.	5. Cinq jours avant la truite sont publiis. *Cinq jours avantque la truite son relâchée.*	5. ห้า วัน ก่อน ที่ ปลาเทราท์ นั้น จะ ถูก ปล่อย *ห้า วัน ก่อน ที่ ปลาเทราท์ นั้น จะ ถูก ปล่อย*

Table 8 Commercial MT System contrasted with QDMT

The third column illustrates the TC in Thai of the first colum from English. The TC which was generated by QDMT is shown numbered and the correct translation is shown directly below in italic font. Each TC is close to or the same as the correct translation.

It is not our intention to compare QDMT to a commercial MT system. Rather we wanted to illustrate how QDMT, with a simple application of constraints and principles, could obtain impressive results, obtaining TCs for subsequent processing. We have designed the application of constraints and use of semantic principles to keep within the spirit of modern unification-based approaches (e.g., HPSG, GPSG, etc.) to language analysis, using appropriate information when needed and subscribing to the general principle of compositionality of meaning.

Also, in fairness to the commercial MT system, it was used to translate English to French and French to English. QDMT generates translations one way at present, from English to Thai. A future MT system based on the GRMT architecture would be greatly enhanced if the translation were performed both ways as well.

6. Concluding Remarks

GRMT is aimed at performing an accurate translation which is focused mainly on keeping the meaning of the input sentence. The central idea behind the GRMT approach is to generate a TC by *QDMT* and then investigate the accuracy of that TC against some acceptable threshold by *conceptual comparison* and *repair and iterate* phases. The speed and complexity of these latter phases directly depend on the output of QDMT. Therefore, QDMT has been designed to generate the TC quickly and as accurately as possible with only a few simple constraints. To reach this TC, QDMT selects the appropriate TL word for each SL word by considering the different characteristics between languages, semantics and word connections without performing any any sophisticated analysis and then rearranges all selected words according to the ordering rules which are learned from a number of example sentences against the Thai grammar.

To demonstrate the capabilities of the QDMT, an initial version of the translation from English to Thai has been developed and run under SICStus Prolog 2.1, on SUN workstation (because the existing Thai keyboard map works properly only on SUN workstation). The QDMT was tested to generate the TC for a number of sentences by using the developed dictionary which contains 80 English words and 150 Thai words. It can generate the most appropriate TC for the input sentence quickly and with relative accuracy. In many cases the translation is accurate without the need for subsequent processing.

Obviously much work remains to be done at this point. QDMT requires some further extensions to enlarge the class of sentences it can handle easily. We are currently pursuing some ideas to extend QDMT. Most of the future development will go into designing the conceptual comparison mechanisms. We first must develop the HPSG parser for Thai, having already available an acceptable HPSG parser for English. Integrating the *compare and iterate* steps into GRMT should present a plethora of as yet unknown problems we eagerly await to tackle. The initial experiments with QDMT have exceeded our expectations; we are optimistic as we begin to analyze the other phases as well.

References:

[Brown, et al., 1992] Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lafferty J.D. and Mercer R.L. (1992) *Analysis, Statistical Transfer, and Synthesis in Machine Translation*, In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Montreal, p. 83-100.

[CICC, 1995] CICC, (1995) Technical Report: Thai Basic Dictionary, Machine Translation System Laboratory, Center of the International Cooperation for Computerization, Tokyo.

[Jone, 1996] Jones, D. (1996) *Analogical Natural Language Processing*, UCL Press Limited, London.

[Longman, 1992] *Dictionary of Contemporary English*, (1992) Longman, Essex.

[Nirenburg et al.,1992] Nirenburg, S, Carbonell J., Tomita M., and Goodman, K. (1992) *Machine Translation: A Knowledge-Based Approach*, Morgan Kaufmann, San Mateo, CA.

[Pollard and Sag, 1987] Pollard, C. and Sag, I. A. (1987) *Information-Based Syntax and Semantics* Lecture Notes No.13, Stanford, Calif: CSLI Publication.

[Sornlertlamvanish et al., 1994] Sornlertlamvanish, V., Pantachat, W. and Meknavin, S. (1994) *Classifier Assignment by Corpus-Based Approach*. In Proceedings of COLING 94: The 15th International Conference on Computational Linguistics, Kyoto, Japan, Vol 1, p556-561.

[Tantisawetrat et al., 1991] Tantisawetrat, N. and Sirinaovakul, B. (1991) *An Electronic Dictionary for Multilingual Machine Translation*. In Proceedings of the Symposium on Natural Language Processing in Thailand, Chulalongkorn Univ., p. 377-402.