[TMI 99: *Proceedings of 8th International Conference on Theoretical and Methodological Issues in Machine Translation*, August 1999, Chester, UK]

Pipelined Multi-Engine Machine Translation: Accomplishment of MATES/CK System

Min Zhang and Key-Sun Choi

KORTERM, Computer Science Department, Korean Advanced Institute of Science and Technology, 373-1 kuSong-dong Yusong-ku Taejon 305-701 Korea
E-mail: {zm, kschoi}@world.kaist.ac.kr

Abstract

In this paper, we propose a new pipelined multi-engine approach to machine translation, which can take advantage of the previously proposed methods, such as rule-based, example-based, pattern-based and statistics-based methods, and eliminate their disadvantages. Some key new techniques in the multi-engine approach, including attribute knowledge classifications, statistical decision-making, pattern transfer, are discussed. MATES/CK, a Chinese-to-Korean Machine Translation system based on the proposed approach, has been developed.

1 Motivation

Many different approaches (Choi *et al.* 1994; Chen & Chen 1995; Su *et al.* 1995; Furuse & Iida 1992; Brown 1996; Brown *et al.* 1993; Yamabana *et al.* 1997; Frederking *et al.* 1994) to machine translation have been advocated these days. But it is generally agreed that no approach, whether rule-based, example-based, pattern-based or statistics-based, is completely adequate in all aspects to the machine translation task. So it is natural to integrate the advantages of these approaches and get rid of their disadvantages in designing a hybrid MT system. Motivated by this consideration, we propose a new hybrid pipelined multi-engine approach to MT. Based on the proposed approach, a Chinese-to-Korean MT system (hereafter, we term it as MATES/CK) has been developed. In the meantime, some key new techniques in the proposed approach, including attribute knowledge classifications, statistical decision-making, pattern transfer and knowledge acquisition, are also proposed and implemented in our MATES/CK system. We will illustrate our approach and the key new techniques with MATES/CK system.

This paper is organized as follows. Section 2 discusses the design philosophy of MATES/CK from the translation engine and translation flow viewpoints. Section 3 discusses some key techniques in MATES/CK system, respectively. Section 4 gives the experiment. Some conclusions are drawn in section 5.

2 The Design Philosophy of MATES/CK — Multi-Engine Model

2.1 Pipelined Multi-Engine MT Model from the Engine Viewpoint

The core idea of MATES/CK system is “pipelined multi-engine”. Each MT engine employs a different MT technology. When using the pipelined multi-engine MT approach, an MT task is divided into many sub-problems and we start up an engine to resolve the corresponding sub-problem that is most suitable for being resolved by the most appropriate engine. According to Frederking et al.’s definition (Frederking et al. 1994), multi-engine machine translation (MEMT) feeds an input text to several MT engines in parallel. But MATES/CK employs different engines serially, not in parallel. So we terms our proposed approach as a pipelined multi-engine approach to distinguish it from Frederking et al.’s definition (Frederking et al. 1994). The pipelined multi-engine MT model here also follows the typical three-phase scheme (analysis/transfer/synthesis) of a conventional transfer-based system.

Rule-based Engine

The rule-based engine is mainly used in the post-processing of Chinese morphological analysis and the pruning processing in the syntactical analysis stage (Zhang & Choi 1999; Zhang 1997). To improve the robustness of the rule-based engine, we propose a linguistic attribute knowledge classification method to quantify the attribute knowledge descriptions slightly, based on which, a new attribute-pruning algorithm is proposed in the Chinese syntactic analysis stage. Further details see section 3.

Statistics-based Engine or Corpus-based Engine

The statistics-based or corpus-based engine is rather encouraging than other engines. We use it in POS tagging, best syntactic tree selection, mapping pattern extraction, and lexical translation. A new probabilistic model was proposed and adopted to select the best syntactic tree from the syntactic tree candidate set (Zhang & Choi 1999). A new lexical selection algorithm was proposed by using Viterbi algorithm and some statistical knowledge (Zhang & Choi 1999).

Pattern-based Engine and Example-based Engine

Patterns usually can capture more sensitive context than rules, for example, a sentence level pattern can describe the whole sentence structural information, but a rule can not. So we use the pattern-based engine in the structural transfer. Our patterns are extracted from examples semi-automatically. We proposed a parameterized pattern-based transfer philosophy (Zhang & Choi 1999). We will elaborate the pattern-based engine in section 3. The example-based engine is partly used in the lexical translation.

2.2 Translation Flow

Figure 1 illustrates the architecture of the pipelined multi-engine model from the translation flow viewpoint, where "PA-Structure Analyzer" is a Chinese predicate-argument (PA) structure analyzer and "P-Bilingual Dictionary" is a bilingual dictionary with the word-aligned translation probabilities. The proposed MT model is described as follows:

- Analysis module is composed of a Chinese morphological analyzer, a parser and a PA detector. The rule- & statistics-based engines are started up in this module. The syntactic parsing includes the construction of syntactic-tree candidate set and the best

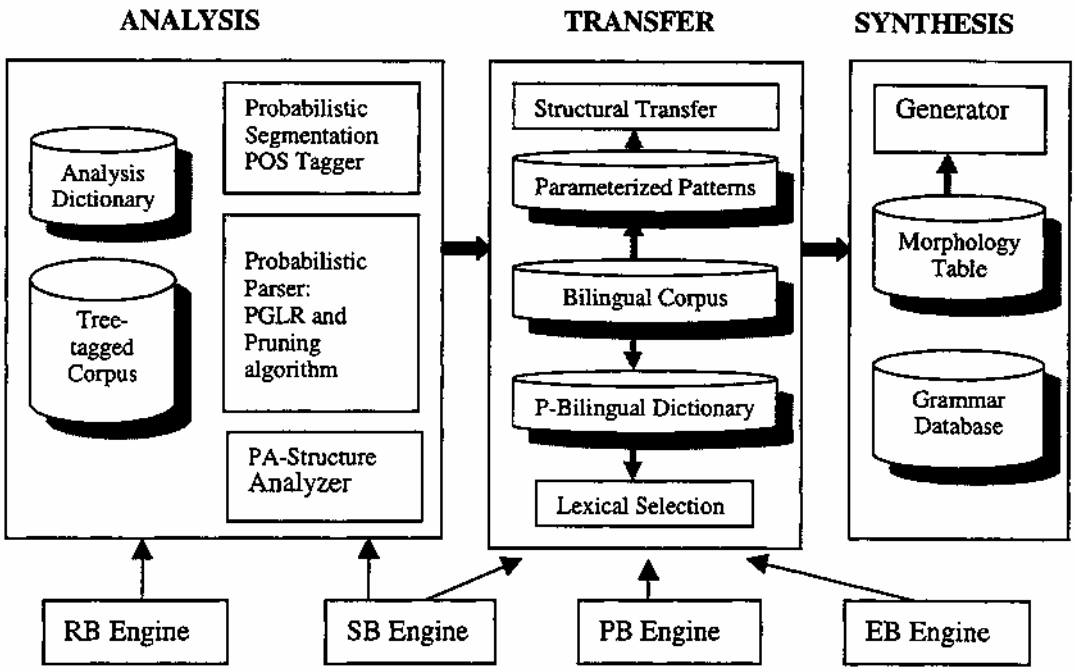

Figure 1. The Architecture of MATES/CK system

tree selection. We use 56 piece of rules to detect the PA structure. Based on the parsing tree and the detailed electronic dictionaries (Yu *et al.* 1998; Mei *et al.* 1985), it is easy to determine the PA structure. An example is listed as follows:

- Chinese input: "你的论文使我对你的工作非常感兴趣。"(Your paper makes me more interested in your works.)

你/pron 的/u 论文/n 使/v 我/pron 对/prep 你/pron 的/u 工作/n 非常/adv 感/v 兴趣/n 。/punct
you GEN paper make me in you GEN works very feel interest .

- Segmentation and POS-tagging:

  where, "adv/GEN/n/pron/v/u/punct" stand for "adverb/genitive/noun/pronoun/ verb/ auxiliary word/punctuation", .

- Tree Candidates constructed by GLR (Generalized Left Reduction) and our pruning algorithm:
  1) [CS[SS[NP 你/pron 的/u 论文/n][VP 使/v 我/pron[VP[PP 对/prep[NP 你/pron 的/u 工作/n]][VP 非常/adv[VP 感/v 兴趣/n]]]]]。/punct]
  2) [CS[SS[NP 你/pron 的/u 论文/n][VP 使/v[SS 我/pron[VP[PP 对/prep[NP 你/pron 的/u 工作/n]][VP{VP 非常/adv 感/v]兴趣/n]]]]]。/ punct}
  3) .....
- The best tree selected by a statistics-based scoring technique:

  [CS[SS[NP1 你/pron1 的/u 论文/n1][VP1 使/v1 我/pron2[VP2[PP 对/prep[NP2

你 /pron3 的/u 工作/n2]][VP3 非常/adv[VP4 感/v2 兴趣/n3]]]]。/punct].[1]

+ PA structure detector: PA(VPl) = "pivotal", PA(VP3) = "collocation"

- Transfer module consists of a lexical selection component and a structural transfer component. The pattern- & statistics-based engines are started up in this module. Structural transfer method is carried out by means of parameterized patterns. Viterbi algorithm is used to carry out the lexical selection module. The following pattern is used to transfer the above parsing tree to Korean structure:

C:C1:[NP]+使+C2:[pron]+对+C3:[NP]+C4:[adv]+感+兴趣+C5: [punct]— >				
	make	about	feel	interest
	PIVOTAL	OBJECT	COLLOCATION	
K:C1:[NP]+은+C2:[pron]+로 하여금 +C3:[NP]+에 대해서+C4:[adv]+흥미를 느끼게 한다+C5:[punct]				
	EUN	LO HAYOGUM	E DAEHAESO	HUNGMIRUL NUKKIGE HANDA
	TOPIC	ROLE (make)	ABOUT	COLLOCATION (be interested in)

Here, in the above diagram, the transfer pattern consists of the Chinese pattern in the first line and the Korean pattern in the fourth line. The second line is the English translation of the Chinese words in the Chinese pattern, and the fifth line is the transliteration of the Korean words in the Korean pattern. The third and sixth lines are the syntactic roles of the Chinese and Korean words in the transfer pattern, respectively.

- Synthesis module consists of a generator and a Korean morphological table. The rule-based engine is triggered in this module. The final Korean translation is:

너의 논문은 나로 하여금 너의 일에 대해서 매우 흥미를 느끼게 한다.
Your paper me make your work about very be interested in
(Your paper makes me more interested in your works.)

3 Some of the Key Techniques in MATES/CK System

As discussed above, the "pipelined multi-engine" model is cooperative of a series of multiple engines that are connected according to their apparent roles in each stage of MATES/CK. In this section, we will discuss some of the new techniques of the proposed multi-engine model in MATES/CK system, including attribute knowledge classification in the RB engine for the analysis module and pattern transfer in the PB engine for the transfer module.

3.1 Attribute Knowledge Classification and Attribute Pruning Algorithm

As above-discussed, GLR algorithm (M. Tomita *ed.* 1991) and attribute-pruning[2] algorithm are used to construct the syntactic tree candidate set in the analysis module. A parsing rule is a CFG-type rule, where several pieces of linguistic attribute knowledge can be attached as

[1] "CS" and "SS" mean complete sentence and simple sentence, respectively. "NP1" (你/pron 的/u 论文/n, your papers) is the TOPIC, "VP1"(使/v+我/pron +VP2, make sb. do sth.) is a typical Chinese PIVOTAL structure, "VP3" (对/prep+ NP2+感/v+兴趣/n, be interested in NP2) is a COLLOCATION, so in the PA structure detecting PA(VPl) = "pivotal" and PA(VP3) = "collocation", "非常/adv" (very much) modifies "VP4" as an adverbial.

[2] Here attribute knowledge includes the lexical, syntactic and semantic knowledge of each word pre-defined in electronic dictionary (Yu *et al* 1998; Mei *et al.* 1985).

matching conditions for pruning out the incorrect branches. The pruning error by the improper attribute knowledge descriptions is the big problem for the general attribute-pruning algorithm. To attack this problem, we propose an attribute knowledge classification method. All the attribute knowledge descriptions attached to the CFG rule are divided into four classifications from two dimensions as follows:

Definition 1: "Strongly-restricted" and "Weakly-restricted" attribute knowledge

If a piece of attribute knowledge with a CFG-rule can describe a certain of natural language phenomenon exactly and completely, we define the corresponding attribute knowledge in this CFG-rule as "strongly-restricted" attribute knowledge (briefly, ***"SR"),*** otherwise the corresponding attribute knowledge in this CFG-rule is called "weakly-restricted" attribute knowledge (briefly, ***"WR").***

Definition 2: "Positive" and "Negative" attribute knowledge

If a CFG-rule is allowed to reduce to a non-terminal symbol while a piece of attribute knowledge attached to the rule is satisfied, we say this piece of attribute knowledge in this CFG-rule is "positive" (briefly, ***"P"***). In contrast, when a piece of attribute knowledge is satisfied, but the CFG-rule is prohibited to carry out a reduce action, we say this attribute knowledge in this CFG-rule is "negative" (briefly, ***"N")***.

The following is a typical parsing rule:

#NP→ adj+n 1524 CenterNode=l
[0:SubClass:l(direct modification) ***WR P***] [0: Attributive:N ***SR N***] .....

where "NP/adj/n" stands for noun phrase, adjective and noun, respectively. The first line is the CFG-rule itself, where "CenterNode=l" means the central node of this CFG-rule is the second node "n", 1524 is the occurrence frequency of this CFG rule in our training corpus. The second line is two examples of attribute knowledge, where "SubClass" and "Attributive" are two kinds of attribute knowledge of an adjective, which are defined in the electronic dictionary (Yu *et al.* 1998). The first attribute knowledge "0:SubClass:l" is used to judge if the sub-classification of the adjective is class 1 among several adjective subclasses, where the adjective is the first Chinese word of RHS (right-hand side) of the CFG rule. All the adjectives in class 1 may modify a noun directly. The adjective feature "attributive" means that adjective can modify noun without "的(Genitive marker)" between adjective and noun. "Attributive:N" means that the adjective cannot modify noun without "的", namely, [0: Attributive:N ***SR N***] means that if adj+n has no "的" between them and the feature "attributive" of this adjective is 'N', then they can not be reduced to NP. According to Chinese grammar, our statistical results from our corpus reveal that:

- Even though an adjective can modify a noun directly, the adjective and the noun are not always reduced to a noun phrase. So the first attribute knowledge is ***"WR"*** and ***"P"***.
- If an adjective can not modify a noun without "的" between them, then adj+n are strictly prohibited to reduce to a noun phrase directly. So the second attribute knowledge is ***"SR"*** and ***"N".***

In general, if a piece of attribute knowledge with a CFG-rule occurs frequently in a balanced

tree-tagged corpus, then this attribute knowledge is great possible to be "weakly-restricted" attribute knowledge, and vice versa. The classifications depend on both the occurrence frequency of the rule with attribute knowledge in a tree-tagged corpus and the linguist's judgement (Zhang 1997). For the limit of the paper's length, we have to discuss the algorithm to acquire the parsing rules with attribute knowledge classifications in our other paper (Zhang & Choi 1999A). Linguist's judgements are necessary in the knowledge acquisition algorithm, but the judgement is not time-consuming and limited within a small scope. In addition, apart from treebank we need not any other specially tagged corpus in our knowledge acquisition algorithm. We have obtained 1174 pieces of parsing rules from our training corpus, including 5964 pairs of CFG-rule and attribute knowledge, in which 2710 pieces is ***"SR"*** and 3254 is ***"WR".***

"Strongly-restricted" means that we can describe a certain of language phenomenon clearly and exactly, so we can use the "strongly-restricted" attribute knowledge without bringing any bias. Algorithm 1 illustrates the construction of the syntactic tree candidate set:

Algorithm 1: Construction of the candidate set of Chinese syntactic trees
Input: Chinese POS-Tagged words
Output: A candidate set of Chinese Tree
Method:

(1) GLR as a basic algorithm. Let α stand for an attribute penalty value of a candidate tree and P_{Tree} stand for the probability of a candidate tree. Their initial values are α =l and $P_{Tree}=1$.

(2) Get an action from LR Table. Every action is associated with an action conditional probability P_A [3].

(3) If the current action is a shift action A_s, then do as a standard GLR algorithm and $P_{Tree} = P_{Tree} * P_A$

(4) If the current action is a reduction action A_R [4], then do as follows:

(4.1) Get a piece of attribute knowledge K_c, and let T_{KC} stand for the corresponding tags of K_c.

(4.2) If K_c == NULL, then $\alpha = a * a_1$, go to (4.5).
// The above line is to calculate attribute penalty value when no any attribute knowledge is satisfied.

(4.3) If K_c is not satisfied with the current input, then go to (4.1).

(4.4) If K_c is satisfied, then do:

- If T_{KC} = ***"SR"*** + ***"P",*** then *go to* (4.5).
- If T_{KC} = ***"SR"***+"***N"***, then *go to* (2). *// Only in this case, pruning action occurs.*
- **If T_{KC} = *"WR"+"P",*** $\alpha = \alpha * \alpha_2$, *go to (4.5). //Calculate penalty value*
- If T_{KC} = ***"WR"*** + ***"N",*** $\alpha = \alpha * \alpha_3$, *go to* (4.5). *//Calculate penalty value*

---

[3] P_A is the action probability in LR table, for whose definition please see Zhang & Choi (1999).

[4] A_R consists of a CFG-rule, several pieces of attribute knowledge description annotated with "strongly-restricted" or "weakly-restricted" and "negative" or "positive" tags, a probability P_A.

(4.5) Execute reduce action AR , $P_{Tree} = P_{Tree} * P_A$, go to (2).

where, three empirical parameters α_1, α_2 and α_3 are assigned to compute the penalty value a, we adjust the value of the three parameters so that the correct tree can be ranked as top as possible. In this paper, α_1 =0.1, α_2 = 0.6 and α_3 =0.3. Please note that, the smaller penalty means that the corresponding tree is less possible to be the correct one.

Algorithm 1 consists of a basic GLR algorithm, a new attribute-pruning algorithm and two scoring functions (calculating α and P_{Tree}). GLR algorithm acts as a basic skeleton parsing algorithm. The attribute-pruning algorithm is used to prune out some of the meaningless candidate trees. The first scoring function is to calculate the attribute penalty value α, and the second one is used to calculate the probability P_{Tree} of each candidate tree. Our pruning algorithm plays an important role in algorithm 2. Only in the second case in step 4.4, when the attribute knowledge is annotated with ***"SR"*** and ***"N",*** we can prune out the meaningless branches. In the other cases, we will give the attribute penalty. This can, not only prune out lots of useless candidates, but also guarantee the correct one reserved. Furthermore the attribute penalty reflects the inexactness of the attribute knowledge description, namely, a can indicate which trees are more possible to become the correct one. So our attribute classification method is an effective way to avoid the pruning errors.

Once the Chinese syntactic tree candidate set is constructed, the statistics-based engine will be started up to select the best syntactic tree from the candidate set. Based on the algorithm 1, we propose and employ an integrated scoring function to select the best tree (Zhang & Choi 1999). The scoring function combines the candidate tree probability P_{Tree} with the penalty value a, which can describe the syntactic tree both quantitatively and qualitatively. For further discussion of the scoring function, please see Zhang & Choi (1999).

3.2 Parameterized-pattern-based Structural Transfer

Structural transfer is carried out by means of the pattern-based engine. A mapping pattern is a typical parameterized bilingual sentence or sub-sentence pair with some parameters, which is formalized as:

$$CST_0 + ... + CST_n | (Ph) \rightarrow KST0 | [t_0] + ... + KSTm | [tm] \quad \{P_{Score}\} \quad (0<m<n)$$

"*CST*" and "*KST"* stand for "Chinese Sub-sTructure" and "Korean Sub-sTructure".

CST_i= | a Chinese word or POS | a sub-classification of POS or word semantic category | phrase tag |

KSTi = | an Korean word or POS | a sub-classification of POS or word semantic category | phrase tag |

"Ph" is a Chinese phrase tag, which means that $"CST_0+...+CST_n"$ should be finally reduced to a phrase *"Ph"*. The integer index t_i attached to KST_i means that $\boldsymbol{CST}_{t_i}$ is transferred to KST_i which is used in lexical selection. P_{Score} is a priority evaluation function, which is defined as follows:

$$P_{Score} = \sum_{i=0}^{n} E(CST_i) \quad (1)$$

$$E(CST_i) = \begin{cases} 0 & CST_i \in \text{phrase tag} \\ 1 & CST_i \in \text{POS tag} \\ 2 & CST_i \in \text{sub-classification of POS or word semantic category} \\ 3 & CST_i \in \text{Chinese word} \end{cases} \quad (2)$$

The following are some typical patterns:

P1. [n|pron]+ 看 见 (see)+[NP]+ 放 在 (put)+[n]+ 上 (on)+[punct] | {***CS***}
→[n|pron]|{0}+?+[NP]|{2}+?+[n]|{4} +위에(on)|{5} 놓여(put)|{3} 있는(have) +것을 (sth.)+ 보았다(see)|{1}+[punct]|{6} {***Score***=11}

P2. [n|pron]+v+SS|{SS}→[n|r]1{0}+?+[SS]|{2}+[v]|{1} {***Scored*** =2}

P3. [v]+[n]|{VP}→[n]|{1}+?+[v]|{0}+? {***Score***=2}

P4. 打(play)+排球(volleyball)|{VP}→배구를(volleyball)|[1]+하다(play)|[0] {***Score***=6}

where "?" means that there should be a Korean morphological change or a postposition or an auxiliary word in this position, but which can not be determined in this pattern currently.

From the pattern definition and formula (1), we can find that our patterns are parameterized by associated with a priority evaluation function P_{Score} and a corresponding position index t_i. We can draw some hidden features:

- Formula (1) is a priority evaluation function, which is used to reduce the conflict among patterns. When a conflict occurs, the preferred one is the pattern whose evaluation value is higher. The idea behind formula (2) is that, the more fine-grained linguistic knowledge a pattern contains or the longer a pattern is, the more preferred a pattern is. According to formula (2), the lexical patterns are the most preferred, for example, pattern P4 is preferred to P3, and P1 is preferred to P2.
- "Ph" defined in a pattern can guarantee that a pattern must be matched with a complete syntactic structure. Pattern is a linear continuous string, but it contains a certain of linguistic information, so we limit that only a complete syntactic structure can be matched with a pattern. "A complete syntactic structure" means a truncation of a sub-tree. Figure 2 illustrates what is a truncation. In figure 2, "T_1=pron+v+num+mea+n" is an invalid truncation, but "T_2=NP+v+PP" is a valid truncation, SS is the root of the truncation T_2. Only when T_2 and SS are matched with

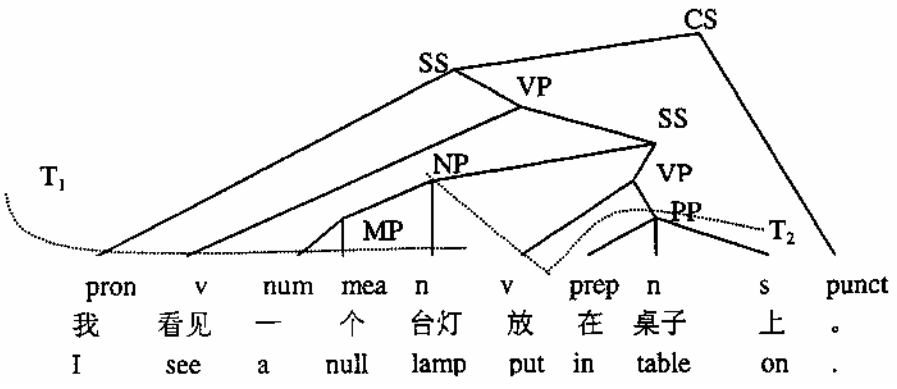

Figure 2. An example of truncation

a mapping pattern successfully, we can say the matching is right. The sentence in figure 3 can be matched with P1 and P2, but P1 is preferred.

- The integer index t_i in a pattern records the important *position* mapping relation between a Chinese word and its possible Korean translation, which is very useful in lexical selection (Zhang & Choi 1999).

Our transfer patterns are extracted semi-automatically from our bilingual corpus. 23,200 mapping patterns are obtained from the corpus.

Once the final syntactic structure and word order of the Korean translation are determined by the mapping patterns, the statistics-based engine will be started up to carry out the lexical transfer processing. We propose a statistics-based lexical transfer method that uses bilingual lexical transfer probability and Korean word co-occurrence statistics as well as Viterbi algorithm. For further discussion, please see Zhang & Choi (1999).

4 Experiment

We built a Chinese-Korean bilingual corpus to train and test MATES/CK system. The corpus contains 115,960 sentences, all of the sentences are Chinese-Korean bilingual pairs, and out of which 61,599 sentences are Chinese-Korean-English trilingual pairs. The average length of the sentences is 13.2 Chinese words per Chinese sentence and 9.2 *eojeois* per Korean sentence. The corpus includes daily sentences and economic texts.

Grammatical Knowledge-Base of Contemporary Chinese (Yu *et al.* 1998) is used as Chinese syntactic knowledge database and <<*TongYiCi CiLin*>> (Mei *et al.* 1985) as a Chinese thesaurus. <<Chinese-Korean Dictionary>> (Hong *et al.* 1989) is used as a basic Chinese-Korean dictionary to tag corpus and get the word translation dictionary for lexical selection.

Total 2100 typical bilingual sentences are selected from our corpus to test MATES/CK system. The test set is also used to train MATES/CK system. 1500 sentences are selected on purposes so that the Chinese syntactic features and the Chinese-Korean bilingual mapping issues can be considered fully in the testing corpus, the other sentences are selected randomly. The average length of the testing sentences is 15.2 Chinese words per sentence.

Based on the above sources, we have got 1174 CFG rules with 2710 ***"SR"*** attribute knowledge and 3254 **"WR"** attribute knowledge as well as a probabilistic LR table for Chinese analysis. We have also obtained 23,200 parameterized mapping patterns for structural transfer and a 4200-entry transfer dictionary for lexical selection[5].

In the analysis module, based on the rule-driven engine, 92.9% syntactic trees are pruned out by our attribute-pruning algorithm[6], at the same time, no any correct syntactic trees are pruned out by mistake. In contrast, if all the ***"WR"*** attribute knowledge is changed to ***"SR",***

[5] All the Chinese words with only one Korean translation are excluded from the 4200-entry transfer dictionary.

[6] This is not surprised, because Chinese language is lack of morphological change and the words order of Chinese sentences are rather free. A test (Zhang 1997) reveals that there will generate 15743 syntactic candidate trees for a simple Chinese sentence "**我们不能学习英语**(we can not learn English)" by using our CFG parsing rules and GLR algorithm without any pruning process. Another example is: considering the CFG rule "SS(Simple Sentence)→n(noun)+adv", "n" and "adv" occur immediately 126,076 times in the corpus, but only in two cases, "n" and "adv" can be reduced to "SS"(one is "**大雨哗哗***(It is raining heavily)",* one is "**大雪纷纷***(It is snowing heavily)"),* so we can find that, in this case, there must be large number of branches should be pruned out.

then there will be 99.1% syntactic trees to be pruned out, but unfortunately 27.2% correct syntactic trees are also pruned out in the meantime. This reveals that the traditional attribute-based method is too rigid to be robust and our attribute knowledge classification method is an effective way to improve the robustness of the attribute-based method.

We give a decision criteria of four levels: best(score=1.0), good(0.6), poor(0.2) and error(0.0) to evaluate the structural transfer and whole translation quality (Choi *et al.* 1994). The final score for evaluation (FSFE) is equal to the arithmetical mean of all the scores:

$$FSFE = \frac{1.0*\#\text{of “best”} + 0.6*\#\text{of “good”} + 0.2*\#\text{of “poor”}}{\text{number of sentences}}$$

Table 1. The FSFE Results

	Word Order	Translation Quality
FSFE	0.873	0.721

From Table 1, the performance of our approach is promising. Please note that the whole translation accuracy should be more than the product of parsing accuracy and transfer accuracy, because in some cases even if the parsing tree is not right, maybe the Korean translation is also right by our transfer patterns. The speed of MATES/CK is very high. It only takes 270 seconds to translate all of the 2100 Chinese sentences with IBM PC 586/400 128M. The main translation errors arise from the analysis and structure transfer of some complex Chinese syntactic or semantic structures and some idiomatic expression translation as well as the Korean generation.

5 Conclusion

Distinguished from the Frederking et al.'s definition (Frederking et al. 1994), we propose a hybrid pipelined multi-engine approach to MT in this paper, based on which MATES/CK system was implemented. We aim at making full use of the different translation engines. According to the proposed MT module, the various problems of a translation task in each phase are decomposed into some sub-problems and each sub-problem is tried to be solved by the most appropriate translation engine. In summary, the proposed approach has the following features and advantages compared with some traditional approach:

- It can integrate the different MT approach naturally, and each MT sub-problem can be resolved by the most appropriate translation.
- Linguistic attribute-knowledge classification method can improve the linguistic knowledge-based methods greatly.

Our future research will be directed towards the construction of large training corpus and exploitation of more powerful hybrid MT language model.

Acknowledgments: We are grateful to Miss Song, Heejung, Ms. Huang, Jinxia, Prof. Wu, Yonghua, Miss Song, Youngmi and Miss Kim, Jihyoun, who are our partners, for their fruitful collaboration and help.

References

Brown, F. Peter, Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer: 1993, *The Mathematics of Statistical Machine Translation: Parameter Estimation,* Computational Linguistics, 19(2), 223-311

Brown, Ralf D.: 1996, *Example-based machine translation in the Pangloss system,* in 16th International Conference on Computational Linguistics: COLING-96, pp.169-174

Chen, Kuang-hua & Hsin-His Chen: 1995, *Machine Translation: An Integrated Approach,* in 6th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95, pp.287-294

Choi, Key-Sun, Seungmi Lee, Hiongun Kim, Cheoljung Kweon & Gilchang Kim: 1994, *An English-to-Korean Machine Translator: MATES/EK,* in 15th International Conference on Computational Linguistics: COLING-94, pp.129-133

Frederking, R., Nirenburg, S., Farwell, D., Helmreich, S., Hovy, E., Knight, K., Beale, S., Domashnev, C., Attardo, D., Grannes, D. and Brown, R.:1994, "Integrating Translations from Multiple Sources within the Pangloss Mark III Machine Translation", in 1st Conference of the Association for MT: AMTA-94

Furuse, Osamu & Iida Hitoshi: 1992, *An example-based method for transfer,* in 4th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-92,pp.l39-150

Hong, Ilsik, Jaeho Jung *et al.:* 1989, *<<Chinese-Korean Dictionary>>,* Institute of national culture of Korean university

M. Tomita, *ed.:199l, Generalized LR Parsing,* Kluwer Academic Publishers

Mei, Jia-jv, YiMing Zhu, Yunqi Gao & Hongxiang Yin: 1985, *Chinese thesaurus: TongYiCi CiLin,* Shanghai Dictionaries Press (in Chinese)

Su, Keh-Yih, Jing-Shin Chang & Yu-Ling Una Hsu: 1995, *A Corpus-based Two-Way Design for Parameterized MT System: Rational, Architecture and Training Issues,* in 6th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95, pp. 334-353

Yamabana, Kiyoshi, Shin-Ichiro Kamei, Kazuniri Muraki, Shinko Tamuba & Kenji Satoh: 1997, *A hybrid approach to interactive machine translation—integrating rule-based, corpus-based, and example-based method,* in 15th International Joint Conference on Artificial Intelligence: IJCAI-97, pp.977-982

Yu, Shiwen, Xuefeng Zhu, Hui Wang & Yungyung Zhang: 1998, *the Grammatical Knowledge-base of Contemporary Chinese—A Complete Specification,* Tsinghua University Press (in Chinese)

Zhang, Min: 1997, *Research on Algorithm of Chinese Treebank Construction Based on Weakly Restricted Stochastic Context-Sensitive Grammars.* Ph.D. dissertation, CS Dept., Harbin Institute of Technology University, P.R.C, Oct. 1997 (in Chinese)

Zhang, Min & Key-Sun Choi: 1999, *Pattern-based and statistics-oriented Chinese-Korean Machine Translation,* in 18th International Conference on Computer Processing of Oriental Languages: ICCPOL'99, pp. 93-98

Zhang, Min & Key-Sun Choi: 1999A, *Attribute-Knowledge Classification and Statistical Decision-Making in Chinese Parsing,* to submit to the 5th Natural Language Processing Pacific Rim Symposium (NLPRS'99)

Filename: Zhang.doc
Directory: G:\MT conferences\TMI 1999
Template: C:\Documents and Settings\John Hutchins\Application Data\Microsoft\Templates\Normal.dot
Title: [TMI 99: Proceedings of 8th International Conference on Theoretical and Methodological Issues
Subject:
Author: John Hutchins
Keywords:
Comments:
Creation Date: 11/15/2006 9:34:00 AM
Change Number: 6
Last Saved On: 11/15/2006 11:06:00 AM
Last Saved By: John Hutchins
Total Editing Time: 78 Minutes
Last Printed On: 11/15/2006 11:06:00 AM
As of Last Complete Printing
Number of Pages: 11
Number of Words: 4,357 (approx.)
Number of Characters: 24,835 (approx.)