A Pipelined Multi-Engine Approach to Chinese-to-Korean Machine Translation: MATES/CK

Min Zhang
KORTERM, CS Department, Korean Advanced Institute of Science and Technology, 373-1 Kusong-Dong Yusong-ku Taejon 305-701 Korea

Key-Sun Choi
KORTERM, CS Department, Korean advanced Institute of Science and Technology, 373-1 Kusong-Dong Yusong-ku Taejon 305-701 Korea

Abstract

This paper presents MATES/CK, a Chinese-to-Korean machine translation system. We introduce the design philosophy, component modules, implementation and some other aspects of MATES/CK system in this paper.

1 Introduction

A Chinese-to-Korean Machine Translation system MATES/CK has been developed as a research prototype and is still under upgrading now in KAIST (Korea Advanced Institute of Science and Technology). Up to now, though many different approaches (Choi et al. 1994; Su et al. 1995; Brown et al. 1996; Frederking et al. 1994) to MT have been advocated, 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. In order to integrate the advantages of these approaches and get rid of their disadvantages in designing a hybrid MT system, we propose a new hybrid pipelined multi-engine approach to MT and apply this approach to our MATES/CK system. We introduce the design philosophy, component modules, implementation and some other aspects of MATES/CK system in this paper.

2 Design Philosophy

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). To improve the robustness of the rule-based engine, we propose a linguistic attribute knowledge classification method and a new attribute-pruning algorithm (Zhang 1997; Zhang & Choi 1999).

Statistics-based Engine

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

Pattern-based engine is used in structural transfer. Our patterns are extracted from examples semi-automatically. We proposed a parameterized pattern-based transfer approach (Zhang & Choi 1999).

3 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:

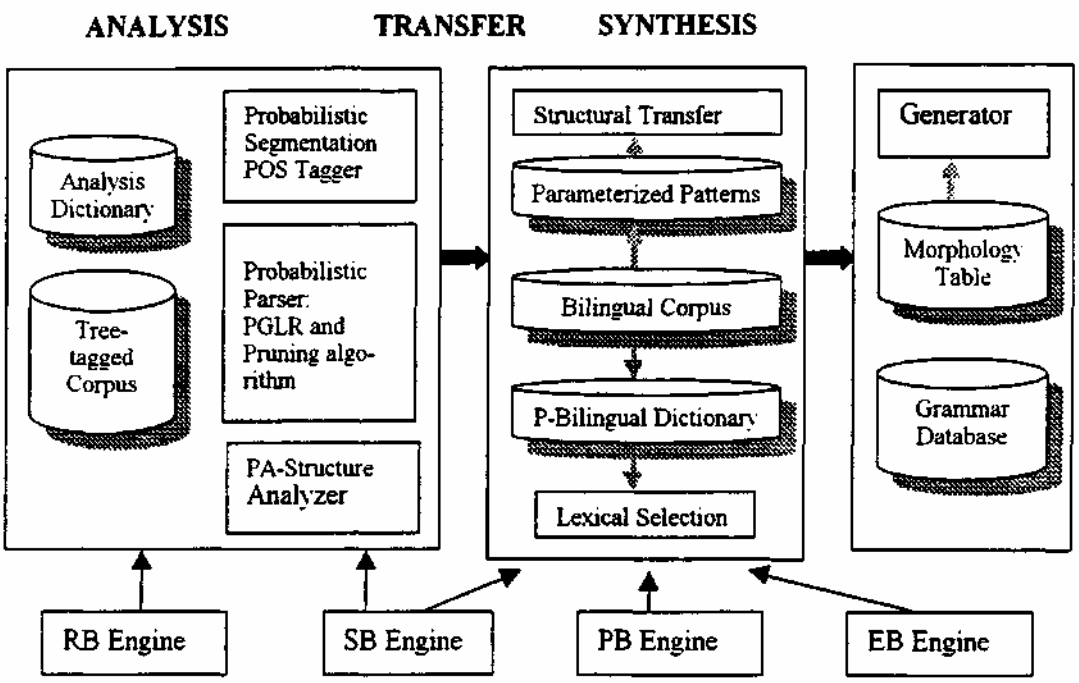

Figure 1. The Architecture of MATES/CK system

- 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 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). it is easy to determine the PA structure. An example is listed as follows:
  1) 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.)
  2) Segmentation and POS-tagging:
  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
  3) 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) ….
  4) The best tree selected by statistics-based 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]
  5) 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:

[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.

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 the module. The final translation is:

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

4 Implementation and System Scale

MATES/CK system was implemented with Visual C++ programming language under Win98 OS in the end of 1998. We adopt the object-oriented internet-based programming design philosophy when developing MATES/CK system.

We built a Chinese-Korean bilingual corpus to train and test the 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 corpus are obtained from the following data sources:

- KAIST corpus (Kim & Choi 1996). 12.000 Chinese-Korean sentences are obtained from this corpus.
- HIT corpus (Zhang 1997). This corpus includes 61,599 Chinese-English sentence pairs; we translated all the sentence pairs from Chinese and English to Korean.
- 23,000 sentence pairs are obtained from the examples sentences of «Chinese-Korean Dictionary» (Hong *et al.* 1989).
- The other sentences are obtained from some newspapers and some books.

The average length of the sentences is 13.2 Chinese words per Chinese sentence and 9.2 *eojeois* per Korean sentence. The domain of this corpus is about the daily sentences and economic domain as well as some news style.

We use Beijing University's Grammatical Knowledge-Base of Contemporary Chinese (Yu *et al.* 1998) as Chinese syntactic knowledge database and *«TongYiCi CiLin»* (Mei *et al.* 1985) as the Chinese thesaurus for describing Chinese word senses «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.

Based on the corpus and the dictionaries, we have got 1120 probabilistic parameterized rules for Chinese segmentation, two probability matrixes and 321 rules for Chinese POS-tagging, 1174 CFG rules with 2710 "strongly-restricted" attribute knowledge and 3254 "weakly-restricted" attribute knowledge (Zhang & Choi 1999) as well as a probabilistic LR table for Chinese analysis. Furthermore, we have also obtained 32,200 parameterized mapping patterns for structural transfer, two probability matrixes and a 4200-entry transfer dictionary for lexical selection[2].

5 Evaluation

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

We test our approach at two aspects. One is the performance of each module and each engine; the other is the whole translation quality.

In the analysis module, based on the rule-driven engine, 92.9% syntactic trees are pruned out by our new attribute-pruning algorithm[3], at the same time, no any correct syntactic trees are pruned out by

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

[3] This is not surprised, because Chinese is lack of morphological change and the words order of Chinese sentences are rather free. A large number of Chinese syntactic trees will be generated by using a pure GLR algorithm. 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 paring rules and GLR algorithm without any pruning process.

mistake. In contrast, if all the "weakly-restricted" attribute knowledge is changed to "strongly-restricted" tag. 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 classification of attribute knowledge is an effective way to improve the robustness of the attribute-based method. (Zhang & Choi 1999)

In the correct syntactic tree-selecting module. based on the statistics-driven engine, out of all the candidate trees, in 81% cases the correct one is the most highly ranked, in 11% cases the correct one is the second ranked. All of the correct trees are ranked within top 10. This result is encouraging.

The evaluation methods of transfer module and the whole translation quality are same. The evaluations of transfer module focus on the word translation (or word sense) and word order. 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 word translation, word order and whole translation quality. respectively (Choi *et al.* 1994). The final score for evaluation (FSFE) is equal to the arithmetical mean of all the scores:[4]

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

Table 1. The FSFE Results

	Word Translation	Word Order	Translation Quality
FSFE	0.912	0.873	0.721

From Table 1. the performance of our multi-engine approach is promising. The result of lexical selection is rather encouraging[5], because no word sense information is employed in our lexical selection algorithm. 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 translation 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.

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.-P., Stephen A.-D.. Vincent J.-D.. and Robert L.-M. (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics. 19(2), 223—311

Choi. K.-S.. Lee. S.-M. Kim. H.-G. Kweon. C.-J. and Kim, G. -C (1994). "An English-to-Korean Machine Translator: MATES/EK". In Proceedings of the 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 Proceedings of the 1st Conference of the Association for MT: AMTA-94

Hong, Ilsik. Jaeho Jung and et al. (1989). «Chinese-Korean Dictionary». Institute of national culture of Korean university

Kim. S.-Y and Choi, K.-S. (1996). "Korean Language Engineering: Current Status of the Information Platform". In Proceeding of the 16th International Conference on Computational Linguistics: COLING-96, pp.1049—1052

Li, J.-J. & Choi K.-S. (1997). "Corpus-Based Chinese-Korean Abstracting Translation System". In Proceedings of the 15th International Joint Conference on Artificial Intelligence: IJCAI-97. pp. 972-976

Mei. J.-J.. Zhu. Y.-Y., Gao Y.-Q., and Yin, H.-X. (1985). «Chinese thesaurus: TongYiCi CiLin». Shanghai Dictionaries Press (in Chinese)

Su, K.-Y.. Chang, J.-S., and Una Hsu, Y.-L. (1995). "A Corpus-based Two-Way Design for Parameterized MT System: Rational. Architecture and Training Issues". In Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation: TMI-95. pp. 334—353

[4] In the denominator in the definition of FSFE, "number of words" is only for evaluating the word translation. "Number of sentences" is only for evaluating the structural transfer and word whole translation quality.

[5] In our previous system, the accuracy of word translation is only 70% (Li & Choi 1997)

Yu, S.-W., Zhu, X.-F., Wang, H., and Zhang. Y.-Y.(1998). "The Grammatical Knowledge-base of Contemporary Chinese—A Complete Specification". Tsinghua University Press (in Chinese)

Zhang, M. (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. M. and Choi, KS. (1999). "Multi-Engine Machine Translation: Accomplishment of MATES/CK System". In Proceedings of 8th Int. Conference on Theoretical and Methodological Issues in Machine Translation: TMI-99