중한 기계번역기 MATES/CK: 파이프라인 번역

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

장민*, 황금하**, 서중원*, 최기선*

한국과학기술원 전산학과 전문용어언어공학연구센터 (*,**)

중국 연변과학기술대학 겸 (**)

{zm,hgh,cwseo,kschoi}@world.kaist.ac.kr

요 약

기계번역기의 방법론인 규칙기반, 예제기반, 패턴기반, 통계기반 각각이 기계번역의 모든 면모를 만족시킬 수 없다는 데에는 이의가 없다. 이러한 여러 방법론의 적절한 융합을 위하여, 이 논문에서는 혼합형 파이프라인 다엔진형 기계번역기로서 중한기계번역기 MATES/CK에 대한 설계 철학, 부분 모듈, 구현 등에 관하여 소개하고자 한다. MATES/CK의 원형시스템 (prototype system)은 이미 구축되었으며 전체 시스템은 여전히 구현 및 보완 중에 있다.

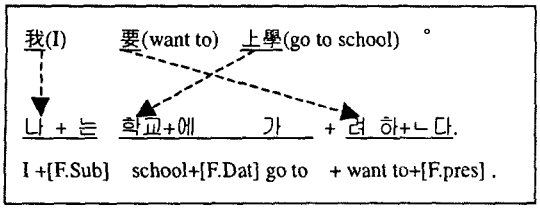
1. 서론

기계번역에 대하여 지금까지 많은 방법론들이 제기되었다 (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). 그러나 기계번역 자체가 많은 복잡한 문제들을 가지고 있는 만큼 어떤 한 가지 방법으로 - 그 방법이 규칙기반이건, 예제기반이건, 패턴기반이건, 아니면 통계기반이건 - 기계번역의 모든 문제를 원만하게 해결할 수는 없다.

중국어와 한국어 간의 기계번역은 두 가지 언어의 서로 상이한 언어적 특성 때문에 기계번역의 여러 처리 단계에서 어려움을 봉착하게 된다. 예를 들면 중국어와 한국어는 그 어순이 완전히 다른데 중국어는 SVO순의 어순을 따르고 한국어는 SOV의 어순을 따른다. 또한 중국어와 한국어는 그 언어적 단위가 상이한 바 1개의 중국어 언어적 단위에 1.9개의 한국어 형태소가 대응하는 것을 볼 수 있다.

이러한 예로 다음 중-한 대응 문장을 보기로 하자: "[중] 我要上學. ⇔ [한] 나는 학교에

가려고 한다. " 이 문장에서의 중한 언어적 단위의 대응 관계는 다음과 같다:



이러한 중한 기계번역의 어려움을 감안하여 본 시스템은 기계번역의 서로 다른 단계에서 서로 다른 방법론을 채택함으로써 각 단계마다 그 단계에 가장 적합한 방법론을 적용하고자 노력하였다. 기존의 다엔진형 기계번역 방법론 (Frederking *et al.* 1994)과는 달리 여러 상이한 엔진을 기계번역의 각 단계에 적용하며, 또한 일부 단계에서는 문제의 효과적인 해결을 위해 두 개 이상 엔진의 혼합 엔진을 사용하기는 하였지만 이 중에서 주를 이루는 것은 한 가지 엔진이라는 점에서 기존의 혼합 방식의 기계번역과 다르다. 이러한 방법을 본 논문에서 파이프라인 다엔진형 기계번역 방법론이라고 규정하였다.

MATES/CK 는 Machine Translation Environment System/Chinese-Korean 의 약칭으로 중한 기계번역 시스템이다.

2. 본 시스템에서 적용하고 있는 엔진들

본 시스템은 규칙기반, 패턴기반, 예제기반 및 통계기반의 엔진을 MATES/CK 시스템의 분석, 변환, 생성 등 여러 다른 단계에 적용하였다. 단순히 여러 엔진을 결합하는 것에 그치지는 것이 아닌 문제의 효과적인 해결을 위하여 본 시스템은 각 엔진의 장점에 따라 효과적인 결합을 도모함으로 각 엔진들의 장점을 취하고 약점을 피하는 효과를 도모하였다.

2.1 확률기반 엔진

확률기반 엔진은 주로 중국어 품사 태깅, 중국어 구문분석 트리에서 최적해 선택 (Zhang 1997), 패턴 추출 및 어휘 변환 단계에서 사용한다.

2.2 규칙기반 엔진

규칙기반 엔진은 주로 중국어 형태소 분석과 구문분석의 가지치기 (pruning) 단계에서 확률기반 엔진과 함께 사용하며 (Zhang & Choi 1999; Zhang 1997) 패턴기반의 변환 과정에서 매칭될 합당한 패턴이 없을 경우 규칙기반 엔진을 적용하여 구조변환을 진행하기도 한다. 규칙기반 엔진의 정확도 향상을 위하여 중국어 구문분석의 가지치기 단계에서 분류된 속성 (Classified Attribute)을 이용한다 (Zhang & Choi 1999).

2.3 패턴기반 엔진

구문 구조가 다른 중국어와 한국어간의 기계번역 시스템의 정확도 향상을 위하여 구조변환 과정에서 패턴기반 엔진을 사용한다. 매 패턴을 그 패턴이 함유하고 있는 정보량 (예를 들면 패턴에 포함된 어휘나 품사정보 등)에 따라 점수를 매겼으며 이러한 점수를 패턴 선택과정에 적용함으로써 정확도를 높일 수 있었다.

2.4 예제기반 엔진

예제기반 엔진은 어휘 변환 과정에서 사용하였는바 관용어적 표현의 번역에 대하여 "작은 패턴"인 예문을 직접 적용한다.

3. 시스템 구성 및 기계번역의 과정

MATES/CK 는 전형적인 변환 방식의 기계번역 시스템으로서 그 구성도는 다음과 같다 (그림 1). 본 장의 나머지 부분에서는 번역의 흐름에 따라 MATES/CK 시스템을 모듈별로 설명하기로 한다.

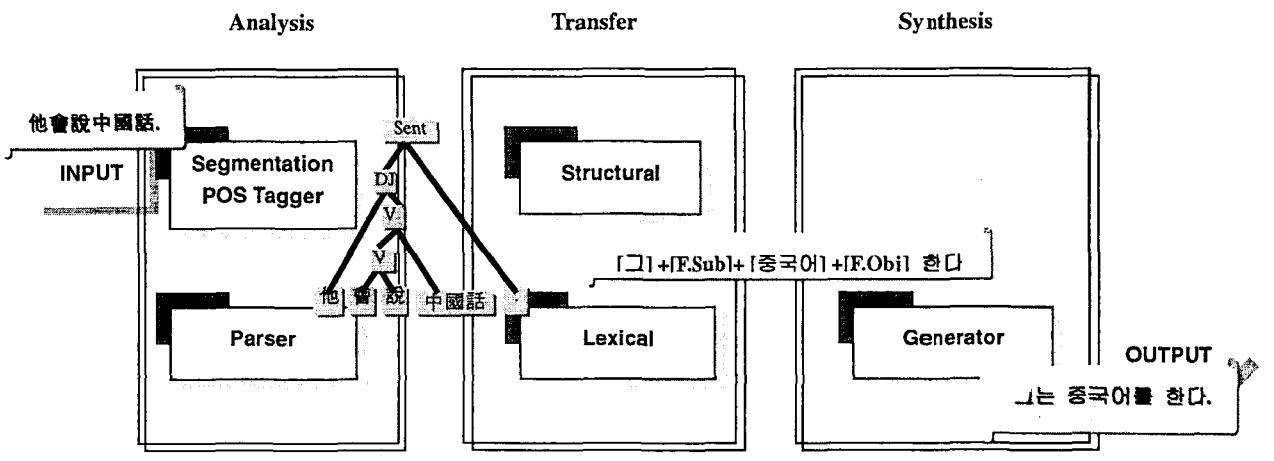


그림 1. MATES/CK 시스템 구성도

3.1 중국어 분석 모듈

중국어 분석사전과 통계정보를 이용하여 FB (Forward and Backward)알고리즘을 적용하여 중국어 단어 분리와 품사 태깅을 진행하였다.

중국어 구문 분석에서는 앞에서 언급하였듯이 규칙 기반 엔진과 통계기반 엔진을 적용하였으며 언어의 분류된 속성지식을 이용하였다. 구문 분석에는 중국어 구문 트리 후보 집합의 구축과 후보 트리 중에서의 최적해 선택 두 단계가 있는데 전 단계에서는 GLR 알고리즘 (Tomita *ed.* 1991) 을 적용하였고 최적해 선택의 가지치기 과정에서는 속성지식에 의한 속성-가지치기 알고리즘 (Attribute-pruning algorithm)을 적용하였다. 속성 지식에는 각 단어별로 기술된 어휘, 구문 및 의미 정보가 포함된 전자 사전이 이용되었다 (Yu, *et al.* 1998; Mei, *et al.* 1985).

일반적인 가지치기 알고리즘의 가장 큰 문제점은 가지치기 과정에서 정확한 트리도 후보 집합에서 가지치기 되어 최적해 선택에서 배제된다는 점이다. 이런 문제를 감안하여 우리는 속성지식을 분류하여 각기 "강한 규제 (Strongly-restricted: SR)"와 "약한 규제 (Weakly-restricted: WR)"로 나누었고 또한 "긍정적 규제 (Positive: P)"와 "부정적 규제 (Negative: N)"로 나누었다. 이러한 속성 지식은 CFG 규칙과 결합하여 적용된다.

속성지식의 분류의 예로 우리는 CFG 규칙 "#NounPhrase → adjective+noun"의 두 가지 속성을 살펴 보기로 한다.

속성(1): "형용사+명사"의 조합이면 NP 로 구문 분석 가능하다.

속성(2): 일부 특정된 형용사 (예: "美麗")는 조사 "的" 없이 직접 명사를 수식할 수 없다.

위의 두 가지 속성에서 속성 (1)은 긍정적 속성이지만 항상 가능한 것이 아니기 때문에 (즉, 두 번째 속성의 제약을 받을 수 있기에) 이는 약한 규제의 속성이다.

그러나 속성 (2)는 그러한 특정된 형용사에 대하여 항상 성립하기 때문에 강한 규제의 속성이며 어떠한 구문구조로 분석될 수 있는 가능성을 완전 배제하는 속성지식이기 때문에 또한 부정적 규제이다.

이러한 두 가지 규제를 다음과 같이 각기 표현하였다.

속성(1): [0:SubClass:1 WR P]

속성(2): [0:attributive:N SR N]

아래의 그림 2 는 CFG 규칙 "#NounPhrase →adjective+noun" 및 분류된 속성 지식을 이용하여 구문분석된 결과에 대하여 가지치기 하는 과정의 예이다.

입력된 중국어 문장 "人(사람)|noun 美麗(아름답다)|adj 花(꽃)|noun 也(도)|adv 美麗(아름답다)|adj." 에 대하여 속성없는 CFG 규칙을 적용한 결과 다음과 같은 두 가지 구문 트리를 얻을 수 있다:

(1) #NP--> [DJ 人+ 美麗] + [DJ 花+[AP 也+美麗]]

(2) #NP--> [DJ[NP 人+[NP 美麗+花]] +[AP 也+美麗]

다음 CFG 규칙 "#NounPhrase → adjective+noun" 의 속성(2)를 적용하면 트리(2)는 가지치기 된다. 즉 주어진 문장의 구문 분석 최적해는 트리(1)인 "#NP--> [DJ 人+ 美麗] + [DJ 花+[AP 也+美麗]]" 이다.

분류 속성을 가진 구문 분석 규칙은 반자동으로 구축되었으며 중국어 구문 분석 알고리즘에서는 GLR 알고리즘을 기본 알고리즘으로 적용하였다 (Zhang &Choi 1999).

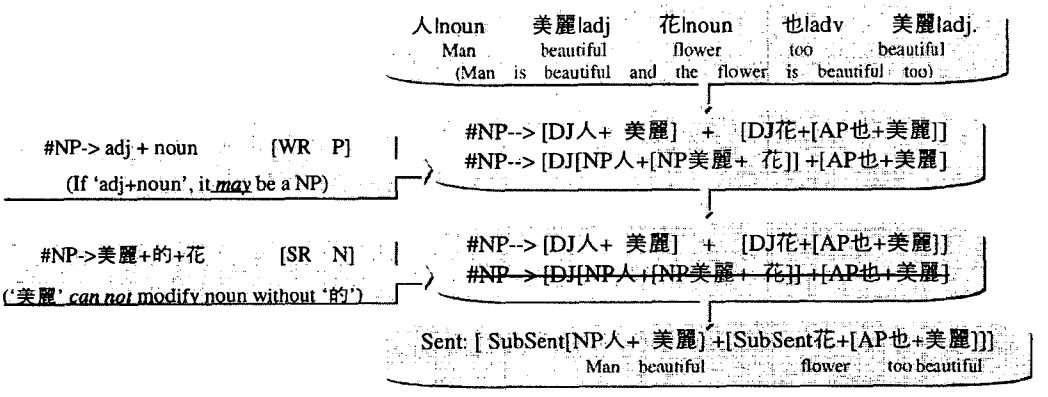


그림 2. 구문분석에서 분류된 속성지식에 의한 가지치기 과정

3.2 변환 모듈

MATES/CK 의 변환 모듈에는 패턴기반의 엔진이 적용되었다. 우선 변환 과정을 그림으로 보면 그림 3 과 같다. 중국어 패턴은 P_{Score} 계산식에 의하여 점수가 매겨진 패턴이며 패턴 점수 계산의 예로 패턴에서 단어로 나타난 "會"와 "說"은 각기 3 점으로, 품사로 나타난 [#1:n/r/np], [#2:n/r/np] 및 [#3:w]는 각기 1 점씩 계산되어 해당 패턴은 총 9 점을 얻는다.

같은 계산식에 의해 입력된 문장과 주어진 패턴 간의 거리가 계산된다. 예제에서 "他/r 會/v 說/v 中國話/n ./w"는 주어진 패턴과의 매칭에서 9점을 얻게 되었고 ("他/r"은 패턴의 [#1:n/r/np]와 품사가 같기에 1점, "會/v"는 단어가 같기에 3점...) 이 점수는 패턴의 점수인 9점과 동점이기에 주어진 패턴은 완전히 매칭되는 것으로 인식되어 대응된 한국어 패턴을 그대로 적용 받게된다. 적용과정에서 원문에서의 단어 및 부호 "他/r", "中國話/n",

"./w"는 각기 패턴에서의 위치 정보 [#1:...], [#2:...], [#3:...]를 적용받아 한국어 패턴에서의 대응된 위치로 들어간다.

이러한 과정을 거쳐 최종 결과인 "[#1:他]+은 +[#2:中國話]+을 한다+[#3:.]"를 얻는다. 이 결과는 어휘 선택 모듈로 들어가 중-한 양국어 사전 및 한국어 코퍼스를 이용한 Viterbi 알고리즘 (Zhang & Choi 1999) 을 이용한 역어 선택 과정을 거쳐 "그+은 +중국어+을 한다+."의 변환단계에서의 최종 결과를 얻게 된다.

적용할 패턴을 찾지 못하는 경우, 규칙 기반의 방식으로 변환을 수행한다. 현재까지 약 2000개의 탑레벨의 변환규칙을 시스템에 적용하였고 이의 하위 규칙까지 약 3000개의 변환 규칙이 시스템에 적용되었다.

대역어 선택에서는 중한 대역어 사전과 한국어 코퍼스를 이용하였으며 현 단계에서는 한국어 어휘간의 공기 정보를 이용하여 단어를 선택하고 있다.

3.3 생성 모듈

한국어 생성에서는 한국어 형태소 테이블과 규칙 기반의 엔진을 적용하여 한국어 기능어를 생성하게 된다. 이 부분에서는 불규칙 동사/형용사의 처리 및 종성 무종성의 처리를 진행하게 된다.

$$P_{Score} = \sum_{i=0}^{n} E(CST_i) \qquad E(CST_i) = \begin{cases} 0 & CST_i \in \text{phrase tag} \\ 1 & CST_i \in \text{POS tag} \\ 2 & CST_i \in \text{sub-classification of POS or word semantic category} \\ 3 & CST_i \in \text{Chinese word} \end{cases}$$

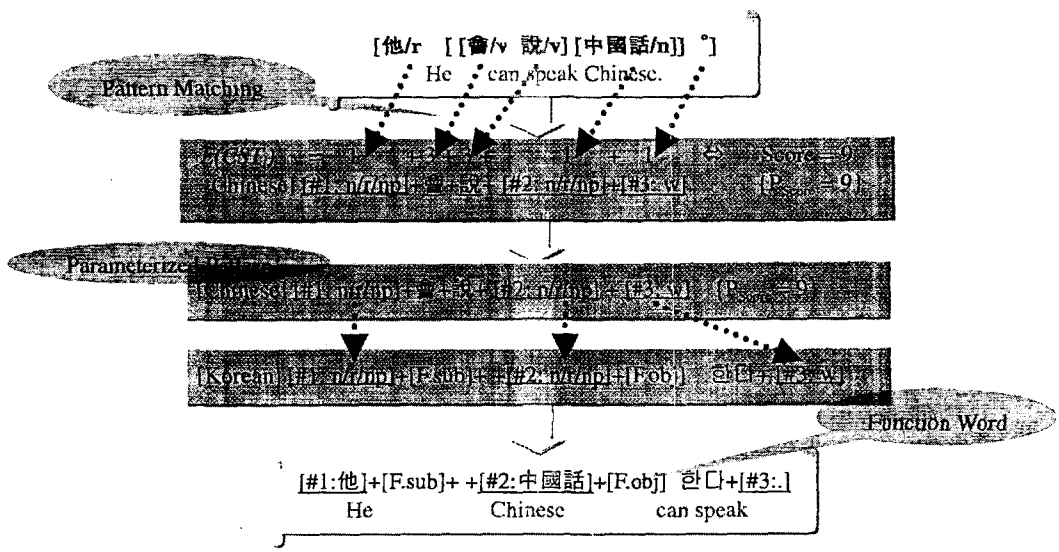


그림 3. 점수가 매겨진 패턴을 이용한 변환과정

4. 실험

우리는 평균 길이가 6.4 분절 단위인 850 문장을 실험대상으로 선택하였다. 이 문장들은 훈련 코퍼스에 포함되어있지 않은 문장이다. 번역 결과를 "1. 정확하게 번역됨", "2. 부자연스럽지만 뜻을 알 수 있음", "3. 추측으로 뜻을 앎", "4. 원문을 참조

해야 뜻을 알 수 있거나 아예 번역문도 아님" 등 4 가지 수준으로 나누었고 여기에서 첫번째와 두 번째 수준을 합격으로 여기고 세 번째와 네 번째 수준을 불합격으로 여겼다.

채점 결과 총 53%의 문장이 합격 (수준 1: 23%; 수준 2: 30%)이고 47%의 문장이 불합격 (수준 3: 27%; 수준 4: 20%)으로 나타났다.

본 연구에서 제안한 분류된 속성에 의한 가지치기 알고리즘에서 92.9%의 구문트리가 가지치기 되었으며 이 중 정확한 트리는 하나도 가지치기 당하지 않았다. 이와 반대로 일반적인 속성지식에 의한 가지치기 알고리즘과 같이 모든 "약한 규제"의 속성을 전부 "강한 규제"의 속성으로 변환해 본 결과 99.1%의 트리가 가지치기 되었지만 이 중 27.2%의 정확한 구문 트리도 함께 가지치기 되는 것을 볼 수 있었다[1]. 이러한 결과로부터 우리가 제안한 분류된 속성 지식을 이용한 가지치기 알고리즘이 효과적임을 볼 수 있었다.

5. 토론

본 연구에서는 혼합형 파이프라인 다엔진형 기계번역 방식을 중한기계번역기 MATES/CK 에 적용하였다. MATES/CK 시스템은 기존 방법론의 기본 사상 (Brown 1996; Chen & Chen 1995; Yamabana *et al.* 1997)을 흡수하였는바, 기계번역의 다양한 문제을 여러 작은 문제로 분리하고 매 문제마다 가장 적합한 엔진을 적용함으로 최적의 번역 결과를 얻기에 노력하였다. 통계기반, 규칙기반, 패턴 및 예제기반의 엔진을 각 엔진의 장점과 특징에 따라 기계번역에서의 분석, 변환 및 생성 단계의 여러 모듈에 각기 적용하였다. 본 논문에서 제안한 분류속성기반의 가지치기 알고리즘은 가지치기 과정에서 대량의 구문 분석 결과 후보로부터 정확한 구문 분석결과를 보호하는데 기여하였다.

한국어에서는 기능어가 구문 구조 및 의미의 전달에서 매우 중요한 역할을 하는 반면 중국어는 기능어가 거의 없기 때문에 생성은 중국어 문장의 의미, 구문구조 등으로부터 얻어야만 한다. 생성에서의 중국어 원문 정보의 부족으로 생성 오류거나 부자연스러운 생성을 얻는 경우가 많은 것을 볼 수 있었고 향후는 보다 다양한 중국어 정보, 예를 들면 존칭, 어휘의 의미정보 등을 얻기 위한 연구를 계속하여야 할 것이다.

중한 번역에서 양국어의 문장 어순이 다른 점을 감안하여 패턴기반의 변환 방식은 번역의 질을 높이는 방법으로 사용되었다. 약 60,000 쌍의 변환 패턴이 구축되었고 현재까지 약 37,000 쌍의 패턴이 시스템에 실제로 적용되었지만 중국어 문장의 어순이 자유롭고 문장 구조가 복잡한 점을 감안하여 패턴기반의 방식에서 얼만큼의 패턴을 적용하여야 시스템이 실용 시스템 (practical system)으로 사용 가능하며, 패턴의 구축에서 어떤 형태의 패턴이 번역의 질적인 향상과 패턴의 보편적 적용이라는 서로 모순되는 문제의 해결에서 더욱 효과적일지는 아직 문제로 남아있다. 향후 최적 패턴 및 규칙 (보다 효과적인 패턴 및 규칙)의 반자동 구축도 연구의 과제로 될 것이다.

6. 참고문헌

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

[1] 이러한 결과는 결코 의외가 아닌바 중국어는 그 언어적 특성에 의하여 문장의 어순이 자유롭기에 GLR 알고리즘만 적용한 구문 분석에서는 대량의 구문 트리 후보를 얻는 경우가 많다. 예를 들면 "我們不能學習英語 (We can not learn English)"라는 문장에 대하여 CFG 규칙과 GLR 알고리즘만 적용한 결과 15743 개의 구문 트리 후보를 얻을 수 있었다 (Zhang 1997).

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.139--150

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

M. Tomita, *ed.*: 1991, *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