에서로/KE:한영 기계 번역 시스템

여상화*,김영길*,최승권*,김태완*,박동인*,서정연**

*기계번역연구실/자연어정보처리연구부/시스템공학 연구소

**서강대학교 전자계산학과

FromTo/KE: A Korean-English Machine Translation

Sanghwa Yuh*, Youngkil Kim, Sungkwon Choi*, Taewan Kim*, Dong-In Park*, Jungyun Seo**

*Machine Translation Lab., Dept. of Natural Language Information Processing, SERI

**Dept. of Computer Science, Sogang University

요약

본 논문에서는 당 연구소 주관으로 연구개발정보센터(KORDIC), 서울대와 공동으로 개발중인 한영 기계번역 시스템, '에서로/KE'의 prototype system을 설명한다. 에서로/KE는 KORDIC에서 한국어 형태소 분석기와 Tagger를 개발하고, 서울대에서 한국어 구문해석기와 한영 변환기를 개발하고, SERI에서 영어 구문 생성기와 영어 형태소 생성기를 개발한다.

한국어 Tagger는 HMM에 기반하여 제작되었으며 sample 200문장에 대해 98.9%의 정확률을 보인다. 한국어 구문 해석기는 의존 문법에 기반하여 CYK 알고리즘을 사용하여 제작되었으며 중의성 해결을 위해 29개의 최적 parse 선택 규칙이 구현되어 있다. 한영 변환기는 collocation과 idiom에 기반하여 한영 변환을 수행한다. 영어 구문 생성기는 Tree 변환 언어인 GWL(Grammar Writing Language)를 사용하여 작성되었으며, 영어 형태소 생성기는 최종적으로 자연스러운 영어 표층문을 생성한다.

에서로/KE는 현재 1차년도 Prototype system이 Unix 환경에서 구현되어 있으며, 현재 각 모듈별 성능 개선과 대량 사전 구축을 통해 상용화될 예정이다.

1. 서론

인터넷의 보급으로 일반인들이 Web을 통해 외국의 문서에 접하는 기회가 많아지면서 외국어를 한국어로 자동번역 하는 S/W에 대한 관심이 높아지고 있다. 이러한 요구에 부응하여 다양한 자동 번역 S/W들이 출시되고 있다. 자동 번역 S/W는 한국어와 일본어간의 구문적 유사성에 근거하여 비교적 높은 번역율을 기대할 수 있는 일한 기계번역 시스템이 주종을 이루고 있다. 영어와 한국어와 같이 다른 어족 간의 자동 번역은 동일한 어족간의 자동 번역에 비해 보다 높은 수준의 번역 기술이 요구된다. 최근 출시되고 있는 여러 종류의 영한 번역 S/W들이 일반 사용자의 기대에 못 미치는 번역 성능을 보이는 것으로 보고되고 있다. 한국어를 외국어로 자동 번역은 한국어 처리에 대한 기반 기술이 취약하여 상용화된 시스템이 드물다. 이러한 S/W로는 시스템 공학 연구소(SERI)와 일본 후지쯔가 공동 개발한 ATLAS II/KJ, 한일 기계번역 시스템과 최근에 발표된 Seoul/Tokyo, 한일 번역 시스템이 있다. 어족이 다른 한영 번역 기계번역 시스템은 Lab Prototype System으로 일부 보고된 것이 있으며[경북대[[서울대], 1994년부터 한국통신의 지원으로 한영 대화체 번역시스템이 서강대와 KAIST를 중심으로 개발되고 있는 대화체 기계번역 시스템이 있다[이현정 97]. 또한, 한국통신에서는 1999년까지 C-STARII Project의 일환으로 한->영/일 자동 통역 시스템(Speech Translation System)을 개발하고 있다[KIM96]. [이현정 97] [KIM96]은 예제 기반의 한영 번역 시스템으로 호텔예약과 항공예약 등과 같은 예약 영역의 대화들을 대상으로 하고 있어 Web문서에 나타나는 일반적인 한국어 문장을 번역하는 데는 부적당하다.

당 연구소는 연구개발정보센터(KORDIC), 서울대와 공동으로 최초의 상용 한영 번역시스템 개발을 목표로 '에서로/KE': 한영 번역시스템을 개발하고 있다.. 에서로/KE는 KORDIC에서 한국어 형태소 분석기와 Tagger를 개발하고, 서울대에서 한국어 구문해석기와 한영 변환기를 개발하고, 시스템공학연구소에서 영어 구문 생성기와 영어 형태소 생성기를 개발한다.

한국어 Tagger는 HMM에 기반하여 제작되었으며

Sample 200 문장에 대해 98.9%의 태깅 정확률을 보인다. 한국어 구문 해석기는 의존 문법에 기반하여 CYK 알고리즘을 사용하여 제작되었으며 중의성 해결을 위해 29 개의 최적 parse 선택 규칙이 구현되어 있다. 한영 변환기는 collocation 과 idiom 에 기반하여 한영 변환을 수행한다. 영어 구문 생성기는 Tree 변환 언어인 GWL(Grammar Writing Language)를 사용하여 작성되었으며, 영어 형태소 생성기는 최종적으로 자연스러운 영어 표층문을 생성한다.

에서로/KE 는 현재 1 차년도 Prototype System 이 Unix 환경에서 구현되어 있으며, 현재 각 모듈별 성능 개선과 대량 사전 구축을 통해 한영 Web 번역기로 개발 중이다.

2. 시스템 구성

에서로/KE 는 변환방식의 기계번역시스템으로 한국어 형태소 분석기 및 Tagger, 한국어 파서, 한영 변환기, 영어 생성기로 이루어져 있다. 그림 2.1 은 **에서로/KE**, 한영번역시스템의 구성도이다.

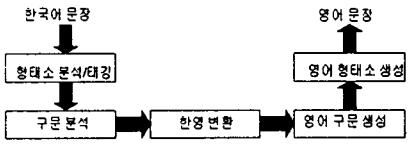

그림 2.1 한영 번역 시스템의 구성도

2.1 한국어 형태소 분석

한국어 형태소 분석기는 지식 확장이 용이하도록 규칙에 기반하여 제작되었다. 규칙 기반 방법의 단점인 처리 속도를 개선하기 위하여 사전은 In-memory Trie 구조를 이용하고 있다[연구개 96]. 현재, 33 개의 어절 구성 규칙과 105 개의 형태 음운 규칙이 작성되어 있으며, 분석에 사용되는 품사 집합은 우리말 정보처리 규격 심포지움에서 제안한 품사 집합에 기반하였다[시스템 97].

형태소 분석의 후처리기로서 HMM 에 기반한 통계적 태거를 두어 최적해를 출력한다[연구개 96]. 태거의 결과는 Viterbi 알고리즘에 의한 최적해 또는 전향(Forward)확률과 후향(Backward)확률을 이용한 최적해를 출력한다.

2.2 한국어 구문 분석기

한국어 구문 분석기는 의존 문법에 기반하여 CYK 알고리즘을 이용하여 의존 파싱을 수행한다. 파서의 전처리 단계에서는 숙어(Idiom)을 인식하여 구문 분석시 발생하는 모호성을 줄이고, 굳어진(frozen) 한국어 쓰임새, 즉 숙어에 대하여 이에 대응하는 영어 번역을 정확히 결정시킨다.

의존 규칙은 이진 규칙(Binary Rule)로 되어 있으며 별도의 규칙 컴파일러에 의해 C 언어로 컴파일된다. 현재 기술된 문법 규칙의 수는 모두 96 개이고 13 개의 문법 관계를 사용한다..

파서의 후처리기로 최적 파스 선택기가 최적 파스를 결정한다. 트리 선택을 위한 휴리스틱 규칙은 현재 29 개가 기술되어 있으며 순차적으로 적용하여 올바른 트리를 선택한다.

2.3 한-영 변환기

구문 분석기의 결과로 최적 파스가 선택되면 변환기는 이를 입력 받아 한국어 트리에 영어 대역어를 붙이고 영어 의존 구조를 만들어 낸다. 영어 의존구조는 기본적으로 한국어 의존 구조와 기능적으로 동일하며 트리의 각 단말 노드는 한국어에 대응하는 영어 어휘를 지닌다. 한국어의 숙어에 대한 영어 노드는 하나의 노드로 처리함으로써 복잡한 한국어 파스 트리에 대해서는 비교적 간단한 형태의 영어 구조를 생성해 낸다. 변환 과정은 숙어 번역, 기본 번역, 조사-어미 번역, 동음이의어 처리, 구조 변환의 5 가지 단계로 이루어져 있다.

변환 사전의 기술 형식은 다음과 같다.

```
("Keyword"
  (1 "default")     ;; Defalt
  (2                ;; Idiom
    ("Korean_idiom1" => "English_idiom1")
    ("Korean_idiom2" => "English_idiom2"))
  (3                ;; Collocation
    (SUB
    (("English_verb")
     ("Korean_noun" => "English_noun"))))
```

실제로 한국어 단어 "먹다"에 대한 변환 사전의 기술 예는 그림 2.2 와 같다.

```
("먹"
  (1 "eat:v")       ;; Defalt
  (2                ;; Idiom
    ("점심!을:n" "=!다" => "have:v" "lunch:n,OBJ"))
  (3
    (OBJ
    (("drink:v")
     ("물" => "water:n"))))
```

그림 2.2 변환 사전의 기술 예

2.4 영어 생성기

영어 생성기는 변환기의 결과인 영어 의존 구조를 입력받아 영어 구구조(Phrase Structure)를 생성하는 '구문 생성기'와 영어 구구조로부터 자연스러운 영어 문장을 생성하는 '형태소 생성기' 로 이루어져 있다.

영어 구문 생성기는 Tree 변환 언어인 GWL(Grammar Writing Language)[한국과 92]를 이용하여 문장 기호 생성, 주어 생성, 어순 조정, 절 생성, 구 생성, 형태소 생성 기능을 수행한다. 특히, 주어 생성의 경우, 영어는 항상 문장의 주어를 필요로 하므로 주절이나 종속절의 주어를 복원해 주어야 한다. 주어 생성 알고리즘은 다음과 같다.

Algorithm *GenerateSUBJECT*()
If (주절의 주어가 없는 경우)
  if(종속절이 없을 경우)
    주절의 주어를 문장속의 문법기능어에 따라 생성;
  else 종속절의 주어를 주절의 주어로 복사;
}
else(종속절의 주어가 없는 경우)
  주절의 주어를 종속절의 주어로 복사;

예를 들어, "나는 살아 있는 동안은 일하고 싶다."의 경우 "살아 있는 동안은"의 생략된 주어가 주절의 주어인 "나"와 동일하게 생성된다.

영어 형태소 생성기는 영어 구구조를 Depth-First 로 운행하면서 단말 노드의 영어 대역어와 구문 정보를 이용하여 시제, 수, 활용처리를 수행한다.

2.5 번역 Workbench

한영 번역 Workbench 에서는 번역 시스템에서 사용하는 해석 사전과 변환 사전 이외에도 다양한 사전(국어, 한영, 영한, 영영한, 전문용어 사전)과 35,776 문장의 대규모의 번역 용례 검색 기능을 제공한다. 이들 사전과 영한 대역 코퍼스의 효율적, 표준적 관리와 재사용성을 위하여 ISO-8879 SGML(Standard Generalized Markup Language)를 사전 및 Corpus 기술 언어로 채택하고 SGML 에 기반한 표준 텍스트/사전 관리기인 TDMS(Text and Dictionary Management System)을 이용한다. TDMS 는 SGML 로 코딩된 사전과 대역 코퍼스를 DBMS 에 기반하여 관리한다[이재성 96][최병진 97].

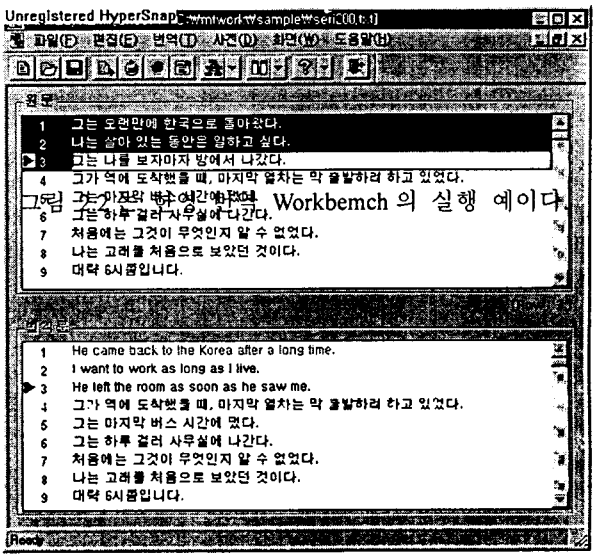

그림 2.3 는 한영 번역 Workbench 의 실행 예이다.

그림 2.3 는 한영 번역 Workbench

3. 실험 및 평가

에서로/KE Prototype System 의 개발을 위해 한국어에서 나타나는 다양한 언어 현상이 반영된 Text 200 문장을 수집하여 Test Set 으로 이용하였다. 번역 문의 평가는 표 3.1 과 같은 기준으로 이루어 졌다.

표 3.1 번역문 평가 기준표

점수	기준
4 (Perfect)	-문장의 의미가 명확 -개별 단어의 번역도 정확.
3 (Good)	-문장의 의미는 대체로 명확 -개별 단어의 번역 오류가 일부 존재(문장의 단어수의 20% 이내)
2 (OK)	-문장의 의미는 몇 번 읽어야 파악됨 -개별 단어의 번역 오류가 일부 존재(문장의 단어수의 30% 이내)
1 (Poor)	-문장의 의미는 추측을 통해 이해됨
0 (Fail)	-여러 번 읽어도 텍스트의 의미를 알 수 없음(non-sense) -번역 실패

표 3.1 은 200 문장에 대한 실험 결과이다. 개발 Platform 은 sun ULTRA 1 170(128MB)이다. 번역문의 일부를 부록 2 에 실었다.

표 3.1 번역 실험 결과

항목	결과
File Size	17,098 Byte(16.7KB)
문장 수	200 문장
전체 어절 수	1,218 어절
문장당 평균 어절수	6.1 어절
문장당 평균 음절 수	19.7 음절
어절의 평균 길이	3.2 음절
번역 속도	Usr: 6.9 Sys:7.1
4 (Perfect)	31 (16%)
3 (Good)	148 (74%)
2 (OK)	11 (5.6%)
1 (Poor)	9 (4.5%)
0 (Fail)	1 (0.5%)

Test Set 의 200 문장에 대한 번역 실험 결과 영어 형태소 생성에 있어서 관사처리, 주어 동사의 성, 수, 시제 일치 등에서 약간의 문제점이 있음에도 불구하고 의미의 전달에는 문제점이 발견되지 않았다. 전체 200 문장에 대한 평균 평가 점수는 3.0((4*31+3*148+2*11+1*9+0)/200)으로 일부의 단어에서 오류가 있지만 문장의 의미 전달에는 문제가 없는 수준이었다.

4. 결론 및 향후 과제

에서로/KE 한영 기계번역 시스템은 현재 1 차년도에 도에는 Sample 200 문장에 대해 번역 실험을 마치고 문제점을 파악하여 개선중이다. 현재 진행 중인 몇 가지 개선 작업들은 다음과 같다,

- 200 만 어절의 Tagged Corpus 를 이용한 품사 Tagging 모듈 재학습
- Tagger 결과로 하나 이상의 n 개의 해석을 출력할 수 있도록 수정
- 파서의 의존 규칙과 트리 선택 규칙을 대규모 Text 에 대해 실험으로 확장 및 보완
- 변환 사전(Idiom 포함)의 확장
- 영어 형태소 생성기 수정/보완

위의 문제점들을 개선하는 작업이 현재 진행 중이며 2 차 년도까지의 결과물을 기업체에 기술 이전하여 상용화할 예정이다.

참고 문헌

[KIM96] Woosung Kim , at al., "A Multi-lingual Speech Translation System for Hotel Reservation," in the Proceeding of '96 Korea-China Joint Symposium on Oriental Language Computing, pp. 25-30, 1996

[서울대 96] 서울대학교, 한영 기계번역을 위한 한국어 구문 분석과 변환에 관한 연구, 시스템공학연구소, 1996

[시스템 97] 시스템공학연구소, 우리말 정보 처리 규격화를 위한 심포지움, 시스템공학연구소, 1996

[연구개 96] 연구개발정보센터, 기계번역용 번역단위 인식 시스템 개발에 관한 연구, 시스템공학연구소, 1996

[이재성 96] 이재성, 최병진, 이운재, 최기선"텍스트 및 전자사전 관리시스템의 설계," 제 8 회 한글 및 한국어 정보처리 학술대회, pp. 408-414, 1996.

[최병진 97] 최병진, 이재성, 이운재, 최기선, "표준화를 위한 사전의 논리구조," 제 8 회 한글 및 한국어 정보처리 학술대회, pp. 415-423, 1996

[한국과 92] 한국과학기술원, 영한기계번역시스템(III): 문법개발지원환경 및 해석 문법 개발, 과학기술처, 1992

[한국과 96] 한국과학기술원, 한.영 텍스트 번역 기술에 관한 연구, 시스템공학연구소, 1996

부록 1: 번역 수행 과정

Input Sentence : 그가 회복될 때까지는 2, 3 주 걸릴 것이다.

Phase I: 한국어 형태소 분석 및 Tagging

[1.1] (N 그) (P 가) (NIND: NOM PRON) (TAG 0.000000)
[2.1] (V 회복되) (E 을) (VIND: REL INTRAN) (TAG 0.000000)
[3.1] (N 때) (P 까지) (AP 는) (NIND: TOPIC) (TAG 0.000000)
[4.1] (N 2) (NIND: NUM ASC) (PUNC ,) (TAG 0.000000)
[5.1] (N 3) (NIND: NUM ASC) (TAG 0.000000)
[6.1] (N 주) (TAG 0.000000)
[7.1] (V 걸리) (E 을) (VIND: REL INTRAN) (TAG 0.000000)
[8.1] (N 것) (CP 이) (E 다) (NIND: AUXN) (VIND: PRES COPL) (PUNC .) (TAG 0.000000)

Phase 2: 한국어 의존 파싱

(BASE: 것, NOUN N: AUXN V: COPL PRES MODAL, E: 다, RULE: ㄹ)
+ADNCL-> (BASE: 걸리, VERB V: INTRAN REL, E: 을, RULE: 87 ADVPMOD)
+NVMODS-> (BASE: 주, NOUN, RULE: 6)
+NNMODS-> (BASE: 3, NOUN N: ASC NUM, RULE: 6)
+NNMODS-> (BASE: 2, NOUN N: ASC NUM)
+ADVP-> (BASE: 때, NOUN N: TOPIC, CP: 까지, AP: 는, RULE: ㄹ)
+ADNCL-> (BASE: 회복되, VERB V: DOEFORM INTRAN REL, E: 을, RULE: 16)
+SUBJECT-> (BASE: 그, NOUN N: NOM PRON, CP: 가)

Phase 3: 한영 변환

Pass 1: 대역어 결정

HEAD-> (것)/ NOUN/ A:v,MAY
+ADNCL-> (걸리)/ VERB/ take:v

```
        +NVMODS-> (주)/ NOUN/ week:n
            +NNMODS-> (3)/ NOUN
                +NNMODS-> (2)/ NOUN
        +ADVP-> (때)/ NOUN/ until:conj *SUB:SUB A:v
            +ADNCL-> (회복되)/ VERB/ get:v well:b
                +SUBJECT-> (그)/ NOUN/ he:pn
```

Pass 2: 영어 의존 구조 생성

```
HEAD-> take /NPOS = 8
    MODS-> week /NPOS = 6
        NMODS->  /NPOS = 5
            NMODS->  /NPOS = 4
    ADVP-> until  get well /NPOS = 3
        SUBJECT-> he /NPOS = 1
```

Phase 4: 영어 생성

Pass 1: 영어 구구조 생성

```
((dummy T

(TEXT ((SCORE 0) )
    (S ()
        (NP ((CASE NOM) )
            (PRON ((BASE IT) (POS PRONOUN) (CASE NOM) )))
        (VP ()
            (VC ()
                (AUX ((BASE MAY) (POS AUX) (TENSE PRES) ))
                (V ((BASE TAKE) (POS VERB) (TENSE NO) )))
            (NP ()
                (NP ()
                    (N ((BASE NIL) (POS NOUN) ))
                    (COORD ((BASE OR) (POS COUNCT) ))
                    (N ((BASE NIL) (POS NOUN) )))
                (N ((BASE WEEK) (POS NOUN) )))
            (BCL ()
                (CONJUNCT ((BASE UNTIL) (POS CONJUNCT) ))
                (S ()
                    (NP ((CASE NOM) )
                        (PRON ((BASE HE) (POS PRONOUN)
    (CASE NOM) )))
                    (VP ()
                        (VC ()
                            (V ((BASE GET) (POS VERB) (TENSE
    PRES) ))
                            (BP ()
                                (ADVERB ((BASE WELL) (POS
    ADVERB) (TYPE NO) )))))))))
    (END ((BASE PMARK) (POS SYMBOL) )))
)
```

Pass 2: 영어 형태소 생성

```
it may take 2 or 3 week until he gets well .
```

Source Korean: 그가 회복될 때까지는 2, 3 주 걸릴 것이다.
Generated English: it may take 2 or 3 week until he gets well

부록 2: 번역 예

Korean: 건강의 고마움은 그것을 잃을 때까지는 알 수 없다.
English: we can not know thankfulness of health until we loses it .

Korean: 서울에 올라오면 언제나 그는 박씨의 집을 방문한다.
English: if he comes up to the seoul he visits always the house of mr.park .

Korean: 그는 의학을 연구하기 위하여 독일로 갔다.
English: he went to germany to study medicine .

Korean: 그는 한국의 풍속습관을 연구하기 위하여 학생들과 함께 내한하였다.
English: he visited korea with the student to study manners and customs of korea .

Korean: 나는 실패하지 않도록 열심히 공부했다.
English: i studied diligently so that i do not fail .

Korean: 그 사고는 그의 부주의 탓이었다.
English: the accident was due to his carelessness .

Korean: 그는 운이 좋았기 때문이 아니라 노력했기 때문에 성공했다.
English: he succeeded not because he was lucky but because he made the efforts .

Korean: 너는 이제 어른이므로 그런 짓을 해서는 안 된다.
English: you must not do such a thing because you is now adult .