2000년 한국정보처리학회 추계 학술발표논문집 제 7 권 제 2 호

용언구에 기반한 한영 기계번역 시스템 : 'CaptionEye/KE'

서영애, 김영길, 서광준, 최승권
한국전자통신연구원 언어공학연구부
e-mail : yaseo@etri.re.kr, kimyk@etri.re.kr, seokj@etri.re.kr, choisk@etri.re.kr

Korean-to-English Machine Translation System based on Verb-Phrase : 'CaptionEye/KE'

Young-Ae Seo, Young-Kil Kim, Kwang-Jun Seo, Sung-Kwon Choi
Language Engineering Department
Electronics and Telecommunications Research Institute
e-mail : yaseo@etri.re.kr, kimyk@etri.re.kr, seokj@etri.re.kr, choisk@etri.re.kr

요 약

본 논문에서는 ETRI 에서 개발 중인 용언구에 기반한 한영 기계번역 시스템 **CaptionEye/KE** 에 대하여 논술한다. **CaptionEye/KE** 는 대량의 고품질 한-영 양방향 코퍼스로부터 추출된 격틀사전 및 대역패턴, 대역문 연결패턴 등의 언어 지식들을 바탕으로 하여, 한국어의 용언구 단위의 번역을 조합하여 전체 번역을 수행한다. **CaptionEye/KE** 는 변환방식의 기계번역 시스템으로서, 크게 한국어 형태소 분석기, 한국어 구문 분석기, 부분 대역문 연결기, 부분 대역문 생성기, 대역문 선택/정련기, 영어 형태소 생성기로 구성된다. 입력된 한국어 문장에 대해 형태소 분석 및 태깅을 수행한 후, 격틀사전을 이용하여 구문구조를 분석하고 의존 트리를 생성해 낸다. 이렇게 생성된 의존 트리로부터 대역문 연결패턴을 이용하여 용언구들간의 연결에 대한 번역을 수행한 후, 대역패턴을 이용하여 각 용언구들을 번역하고 문장 정련과정을 거쳐 영어 문장을 최종 생성한다.

1. 서론

인터넷 환경의 급속한 성장은 인터넷을 이용하는데 있어 가장 큰 문제점 중 하나인 언어 장벽을 극복하기 위해서 기계번역 시스템들에 대한 사용자들의 관심과 필요성을 증폭시켰다. 그러나, 현재 상용화된 번역 시스템들은 비교적 높은 번역율을 보이는 일한 기계번역 시스템이 주종을 이루고 있으며 한국어를 외국어로 자동 번역하는 시스템의 경우 한국어 처리에 대한 기반 기술이 취약하여 상용화된 시스템이 드문 실정이다.[1] 이와 같이 한국어의 처리가 어려운 이유 중의 하나로 한국어의 비정형성을 들 수 있다. 한국어는 영어 등의 언어에 비해 비교적 문법이 자유롭기 때문에 이를 규칙을 기반으로 하여 번역하고자 한 일부 시스템들의 경우에 시스템의 확장성, 견고성, 신뢰성 등에서 여러 가지 문제점을 보였다.[2,3,4]

이러한 문제점들을 해결하기 위해 본 논문에서 기술하는 한영 기계번역 시스템인 **CaptionEye/Ke** 는 데이타에 기반한 번역 방식을 채택하였다. 본 시스템은 대량의 고품질 한-영 양방향 코퍼스를 이용하여 구축된 언어 지식들을 이용하여 한국어의 용언구를 기본 번역 단위로 하여 한-영 기계번역을 수행한다.

현재 ETRI 에서 개발중인 **CaptionEye/KE** 의 개발 목표는 자막 방송이 지원되는 한국어 방송 뉴스를 영어로 번역하여 영어권 시청자에게 영어 자막으로 제공하는 데 있다.

2. 시스템 구성

CaptionEye/KE 는 변환방식의 기계번역 시스템으로서, 번역하고자 하는 한국어 문장에 대해 형태소 분석 및 태깅을 수행한 후, 의존 규칙과 동사구 패턴, MI 정보 등을 이용하여 구문구조를 분석하여 의존 트리

를 생성해 낸다. 생성된 의존 트리로부터 대역문 연결 패턴을 이용하여 한국어 용언구들간의 연결 관계에 대한 영어 번역 정보를 생성해 낸다. 이렇게 해서 번역될 영어 문장에 대한 영어 단문들간 연결을 결정한 후, 용언구 대역패턴과 명사구 패턴을 이용하여 각 용언구들에 대한 영어 문장을 생성한다. 이 후, 문장 정련과정을 거쳐 최종 영어 문장을 생성함으로써 번역을 완료한다. 그림 1 은 **CaptionEye/KE** 의 시스템 구성도이다.

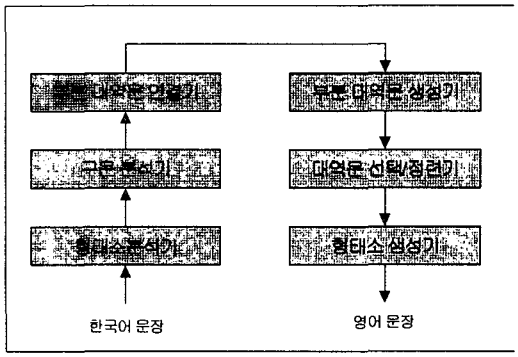

그림 1. 한영 번역 시스템의 구성도

그림 1 에서 보듯이 **CaptionEye/KE** 는 크게 한국어 형태소 분석기, 한국어 구문 분석기, 부분 대역문 연결기, 부분 대역문 생성기, 대역문 선택/정련기, 영어 형태소 생성기로 구성된다. 이하에서는 각 하위 시스템들의 구조 및 처리 방법에 대해 기술한다.

2.1 한국어 형태소 분석기 및 태거

한국어 형태소 분석 시스템은 단계별 기능을 기준으로 전처리기, 형태소 분석기 및 태거로 나눌 수 있다.

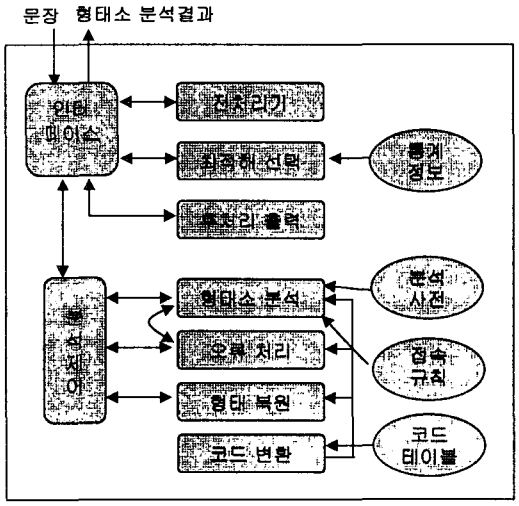

그림 2. 한국어 형태소 분석 시스템 구성도

전처리기는 정규 표현을 이 용하여 입력 문장으로부터 번역 단위인 문장과 어절을 분리하고 한국어 문장에서 나타나는 다양한 표현(예: 영어 인명, 날짜/시간 표현, 단위 표현, 특수 기호 등)에 대응하여 형태소 분석이 용이한 형태로 변환한다.

형태소 분석기는 형태소 분석을 위한 접속정보가 기술된 사전과 좌우 접속 정보표를 이용하여 한국어 어절의 가능한 모든 분석을 밝힌다. 태거는 문맥 확률과 어휘생성확률을 이용하여 한 어절의 가능한 모든 분석 중에서 최적의 분석 후보를 결정한다. 이러한 단계별 기능을 이루기 위해 유기적으로 결합된 시스템의 구성은 그림 2 와 같다.

2.2 한국어 구문 분석기

한국어 의존구조 분석 시스템은 그림 3 과 같이 의존규칙을 참조하여 입력된 문장의 가능한 모든 구조를 분석하는 구조 분석 모듈과, 다양한 정보들을 이용하여 구조 분석 모듈에서 나온 모든 분석 구조 중에서 옳지 않은 분석 구조를 골라내고 미지격을 추정하는 구조선택 모듈로 나눌 수 있다.

구조분석 모듈은 다시 분석 단위인 토큰을 형성하는 부분과 각 토큰 사이의 가능한 의존관계를 조사하는 부분, 그리고 의존관계들을 결합하여 적절한 의존구조를 형성하는 모듈로 나뉘어 진다.

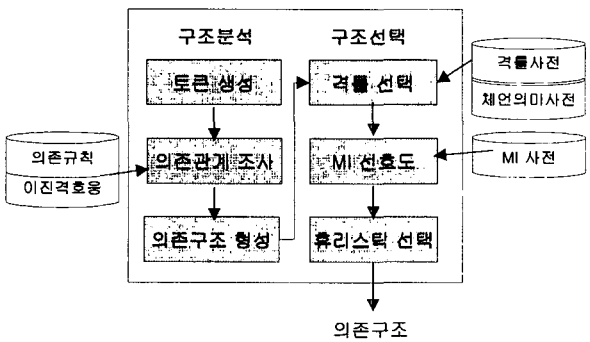

그림 3. 의존구조분석기의 구성도

구조선택 부분은 구축된 용언의 격틀을 이용하여 적절하지 못한 문장 구조를 제거하는 격틀 사용 선택 모듈과 상호정보(MI)를 이용하여 문장 구조의 선호도를 계산하는 상호정보 선호도 계산 모듈, 마지막으로 휴리스틱에 의해 가장 적절한 n 개의 후보를 선택하는 휴리스틱 선택 모듈로 구성된다.

2.3 부분 대역문 연결기

많은 수의 한국어 문장은 하나 이상의 용언구들로 이루어져 있다. 부분 대역문 연결기는 연결 어미 등으로 연결되어 하나 이상의 용언구로 구성된 한국어 문장에 대하여 각 한국어 용언구에 대응하는 영어 번역 결과를 연결하는데 필요한 연결 정보를 제공한다.

CaptionEye/KE 에서는 연결 어미 이외에 일부 관형어미와 체언형 전성어미도 용언구들간의 연결에 사용되는 어미로 간주하여 처리한다.

부분 대역문 연결기는 번역되어 생성될 영어 단문들의 연결을 위한 접속사 정보 및 영어 단문들간의 어순 정보, 형태 정보를 제공해 줄 뿐 아니라, 용언구 고정표현에 대한 번역 정보도 제공해 준다.

용언구 고정표현이란 둘 이상의 용언구로 분산되어 기술되는 고정표현으로, 각각의 용언구에 대하여 대응되는 영어 용언구로 번역한 결과의 결합에 의한 번역이 아니라 그와 다르게 번역되는 모든 경우를 일컫는다.

예를 들어 "이 길로 가면 서울이 나온다"와 같은 문장은 "이 길로 가면"과 "서울이 나온다"의 2 개의 용언구로 이루어져 있으나 번역된 영어 문장에서는 "this load leads to Seoul"과 같이 하나의 용언구로 이루어져 용언구간의 1:1 번역이 이루어지지 않는다. 이와 같은 경우, 이를 용언구 고정표현으로 간주하고, 부분 대역문 연결기는 이러한 형태의 용언구 고정표현의 번역을 위한 모든 정보를 제공해 준다.

이를 번역하기 위해 부분 대역문 연결패턴이라 불리는 패턴을 사용한다.

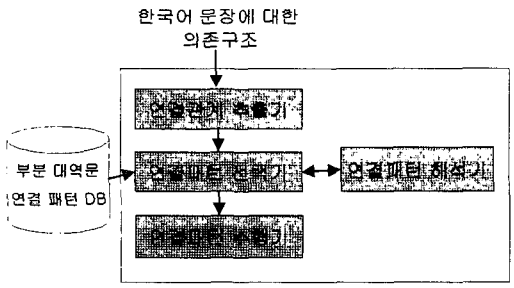

그림 4. 부분 대역문 연결기의 구성도

그림 4 는 부분 대역문 연결기의 시스템 구성도를 나타낸다. 부분 대역문 연결기는 크게 연결관계 추출기, 연결패턴 해석기, 연결패턴 선택기, 연결패턴 수행기의 4 개의 서브 모듈로 구성된다.

연결관계 추출기는 입력된 한국어 구문구조에 대해서 연결관계가 성립 가능한 용언구들간의 모든 연결 관계를 추출한다. 연결패턴 선택기는 연결패턴 DB 를 탐색하여 추출된 연결관계에 대한 연결패턴이 있는지를 검사하고, 연결패턴이 있을 경우는 연결패턴 해석기를 수행하여 현재 입력문의 조건과 연결패턴 내의 제약 조건이 가장 많이 일치하는 연결 패턴을 선택한다. 그 후, 선택된 연결패턴 중 입력문장을 가장 잘 연결할 수 있는 연결패턴의 집합을 선택한다. 연결 패턴 수행기는 선택된 연결패턴 집합내의 패턴들을 이용하여 입력문장에 대하여 영어 대역문장의 순서 결정, 부분 대역문 연결 정보 및 부분 대역문 생성 정보를 제공한다.

2.4 부분 대역문 생성기

부분 대역문 생성기는 한국어 입력문장에서 각 용언구들에 영어 대역문을 생성하는 역할을 한다. 이를 위해 부분 대역문 생성기는 부분 대역문 연결기에서 제공한 생성 정보 외에 각 단문에 대한 영어 대역 정보를 제공하는 부분 대역패턴 정보를 참조한다.

부분 대역패턴은 구문 분석기에서 한국어 문장의 의존 구조를 분석하기 위해서 사용된 격틀 정보와 연결되어 있다. 즉, 구문 분석기에서 각 용언구를 분석하는 데 사용된 격틀 정보의 인덱스를 이용하면 해당 용언구에 대한 대역문을 생성하기 위한 부분 대역패턴을 찾을 수 있다.

또한, 부분 대역문 생성기는 명사구의 번역을 위해 명사구 대역패턴을 이용하고, 부사 처리 규칙을 이용하여 부사들간의 어순을 결정한다.

2.5 대역문 선택/정련기

CaptionEye/KE 의 입력으로 들어온 하나의 한국어 문장에 대하여 한국어 구문 분석기는 n 개의 의존구조 후보를 생성해 낸다. 부분 대역문 연결기와 부분 대역문 생성기는 이들 각각에 대해 해당하는 작업을 수행함으로써, n 개의 문장 구조 후보에 대한 번역 정보를 생성해 낸다. 대역문 선택/정련기는 이들 n 개의 번역 후보에 대해 가장 적합한 하나의 번역 구조를 선택하고, 선택된 영어 번역 정보에 대하여 영어 문장 정련 단계를 거쳐 최종적으로 생성될 영어 문장에 대한 모든 정보를 영어 형태소 생성기에 제공한다.

대역문 선택/정련기는 크게 대역문 선택기, 관계사절 처리기, 대역문 생성 정보 추출기 및 대역문 정련기로 구성되어 있다.

각 서브모듈은 입력된 한국어 문장에 대해 가장 적합한 영어 번역문 구조 선택, 관계사절 연결을 위한 관계사 결정, 영어 문장 생성에 필요한 생략 성분 복원 및 성, 수, 시제 정보 제공, 반복 주어 대명사화 등과 같은 영어 문장 스타일 정련의 작업을 수행한다.

그림 5 는 대역문 선택/정련기의 시스템 구성도이다.

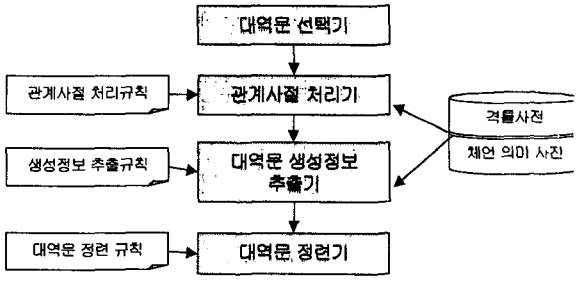

그림 5. 대역문 선택/정련기의 구성도

2.6 영어 형태소 생성기

영어 형태소 생성기는 한영 자동 번역 시스템의

가장 후단에 위치하여 최종적으로 영어 문장을 생성해 주는 기능을 담당한다. 영어 형태소 생성기는 의미, 구문 수준의 영어 생성 결과로부터 영어 문법에 맞는 형태소 수준의 문법처리를 거쳐 자연스러운 표층 문장을 만들어 낸다.

3. 실험 및 평가

CaptionEye/KE 를 평가하기 위해 한국어에서 나타나는 다양한 언어 현상이 반영된 Text 200 문장을 Test Set 으로 이용하였다. 번역 문의 평가는 표 1 과 같은 기준으로 이루어졌다.

표 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 2 는 Test Set 200 문장에 대한 실험 결과이다. 개발 Platform 은 윈도우즈 98 시스템이다.

표 2. 번역 실험 결과

항목	결과
File Size	17,098 Byte(16.7KB)
문장 수	200 문장
전체 어절 수	1,218 어절
문장 당 평균 어절 수	6.1 어절
문장 당 평균 음절 수	19.7 음절
어절의 평균 길이	3.2 음절
4 (Perfect)	41(20.5%)
3 (Good)	154 (77%)
2 (OK)	3 (1.5%)
1 (Poor)	1 (0.5%)
0 (Fail)	1 (0.5%)

Test Set 의 200 문장에 대한 번역 실험 결과, 영어 관사처리 및 생략 주어 복원 및 일부 단어에 대한 대역어 선택 등에서 약간의 문제점들이 있었으나, 평균 평가 점수는 3.16 으로서, 문장 전체의 의미 전달에는 큰 문제가 없었다. 그러나, 현재의 Test Set 은 문장 당 평균 어절 수가 6.1 어절로 비교적 짧은 문장들로서 앞으로 보다 길고 복잡한 문장에 대한 Test 를 수행할 예정이다.

4. 결론 및 향후 과제

CaptionEye/KE 한영 기계번역 시스템은 현재 Prototype 시스템의 개발이 완료된 상태로 문제점을 파악하여 개선 중에 있으며 계속적으로 번역 지식을 구축 중에 있다. 현재 시스템 상에서 가장 크게 보완해야 할 점으로는 다음을 들 수 있다.

- 용언구들간의 올바른 의존 관계 분석
- 대역문 정련 기능의 보완
- 번역 지식 구축의 일관성 유지

앞으로 이상의 문제점을 중심으로 시스템을 계속 보완할 예정이다.

참고문헌

[1]여상화, 김영길, 최승권, 김태완, 박동인, 서정연, "에서로/KE:한영 기계 번역 시스템", 한글 및 한국어 정보처리학회,pp283-287, 1997

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

[3] 서정연, 조정미, 김길창, "한영 대화체 기계번역 시스템," 제 11 회 음성 통신 및 신호처리 워크샵 논문집 제 11 권 1 호, pp.65-70, 1994

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