한영 기계번역을 위한 고정표현 지식베이스 구축

이현호*, 안동언*, 정성종*, 김재훈**, 서영애***, 김영길***
전북대학교 컴퓨터공학과*, 한국해양대학교 기계정보공학부**, 한국전자통신연구원 언어공학연구부***
hhlee@calhp1.chonbuk.ac.kr, {duan,sjchung}@moak.chonbuk.ac.kr,
jhoon@hanara.kmaritime.ac.kr, {yaseo,kimyk}@etri.re.kr

Establishment of Fixed Expression KnowledgeBase for Korean-to-English Machine Translation

Hyun-Ho Lee*, Dong-Un An, Sung-Jong Chung,
Jae-Hoon Kim, Young-Ae Seo, Young-Gil Kim
Dept. of Computer Engineering, Chonbuk Univ.,
Korea Maritime Univ., Electronics and Telecommunications Research Institute

요 약

예제 기반 기계번역 시스템에서 해석의 정확도를 높이기 위해서는 대량의 고품질 대역 코퍼스가 필요하다. 이 대역 코퍼스는 예문을 단순히 나열해 놓은 것이 아니라 일정한 표현 형식에 따라서 기술한 대역 패턴들이다. 본 논문에서는 용언과 필수항인 명사구들로 이루어지는 고정표현을 정의하고 한국어와 영어의 대역 패턴을 기술하여 지식베이스를 구축한다.

빈도수가 높은 용언 5,000개를 중심으로 한영사전에 있는 용례 58,000여 개의 고정표현 지식을 구축하였다. 본 논문에서는 고정표현 지식베이스를 구축하는 과정을 기술하고, 고정표현 지식을 기술하면서 발생하였던 여러 가지 문제점을 예와 같이 기술한다.

1. 서론

70년대와 80년대의 기계번역시스템 개발 초기에는 규칙을 기반으로 하는 번역(RBMT: Rule-Based Machine Translation)이 주종을 이루었으나 확장성, 견고성, 신뢰성 등에서 여러 가지 문제점을 보였다. 그래서, 90년대 들어서면서 번역의 질을 높이기 위한 새로운 방안으로 예제를 기반으로 하는 기계번역(EBMT: Example-Based Machine Translation)이 등장하였다[1].

예제기반 기계번역시스템에서는 대량의 고품질 코퍼스를 필요로 한다. 그러나 이러한 코퍼스 구축의 어려움으로 인하여 예제기반 기계번역의 번역 결과가 좋지 못했다. 이러한 예제기반 기계번역 시스템이 요사이 새롭게 대두된 것은 대량의 단일언어 코퍼스 및 대역 코퍼스의 이용이 가능해졌기 때문이다.

본 논문에서는 용언과 필수항인 명사구들로 이루어지는 고정표현을 정의하고 한국어와 영어의 대역 패턴을 기술하여 지식베이스를 구축한다.

본 논문은 한국전자통신연구원의 지원으로 수행된 "한영 기계번역을 위한 고정 표현 지식 개발"연구의 일부분입니다

기존의 연구 중에서 고정표현과 유사한 것으로 서울대에서 연구한 숙어 표현[2][3]과 KAIST에서 연구한 chunk[4][5]가 있다. 본 논문의 고정표현은 이러한 연구들을 반영하였다.

빈도수가 높은 용언 5,000개를 중심으로 한영사전에 있는 용례 58,000여 개의 고정표현 지식을 구축하였다. 본 논문에서는 고정표현 지식베이스 구축 과정을 기술하고, 고정표현 지식을 기술하면서 발생하였던 여러 가지 문제점을 예와 같이 기술한다.

2. 고정 표현 지식 구축

2.1 원시 자료

빈도순으로 선정된 5,000개의 용언에 대한 예문을 사전으로부터 수집한다. 대상 사전은 시사영어사의 엘리트 한영대사전이다.

2.2 작업 도구

고정 표현 지식을 구축하기 위해 전용 도구를 개발하지는 않았다. 다만, 작업 시간을 단축하고 오류를 줄이기

위하여 한국어 용례를 형태소 단위로 분리하기 위해서 기존에 개발된 한국어 형태소분석기와 품사태깅 시스템을 사용하였다.

2.2.1 한국어 형태소 분석기

한국어 형태소 분석기는 오른쪽에서 왼쪽으로 차트 파싱을 하여 형태소를 분리한다. 형태소 분석기는 1) 토큰 분리, 2) 불규칙 처리, 3) 형태소 분리, 4) 형태소 배열규칙 검사, 5) 미등록어 추정의 단계를 거쳐서 형태소를 분리한다[6].

2.2.2 한국어 품사 태깅 시스템

한국어 품사 태깅 시스템은 가중치 망을 이용한 한국어 품사 태깅 방법을 이용하였다[6]

2.3 고정표현 지식 구축

고정표현은 [7]에서 CFG 형태로 정의된 고정표현 지식의 기술 방법과 작성지침에 따라 구축하였다. 작성 형식은 다음과 같다.

#가다
KS: 학교에 가다
ES: go to school/attend school
KP: A=학교!에 가!다
EP: go to A=school
EP; attend A=school

표제어는 #으로 표시한다. 약자들인 KS, ES, KP, EP는 각각 "Korean Sentence", "English Sentence", "Korean Pattern", "English Pattern"이다. 패턴이 여러 개일 때는 ";"로 구분한다.

작성된 고정표현에서 보듯이 용언을 중심으로 기술하며 명사항에 대해서는 "A, B, C, ..."와 같은 메타 어휘를 이용하여 대응되는 한국어와 영어 어휘를 표시하였다.

용언에 대해서는 시제정보(PAST, PRES, ...), 양상정보(PROG, PERF), 화법(CAN, WILL, MUST, ...) 등의 자질정보를 표기하였고 명사에 대해서는 소유격(POSS), 재귀(REFL) 등의 자질을 표기하였다.

빈도수 높은 용언 5,000개에 대해서 용례 58,000여 개의 고정표현 지식을 구축하였다. 각 용언 당 평균적으로 10여개의 용례를 가지고 있었지만 편차가 매우 컸다.

용례에서 중문이나 복문은 2개 이상의 문장으로 분리해 단문으로 만들어 고정 표현 지식을 구축하였다.

3. 고정 표현 지식 작성의 문제점

고정 표현 지식을 작성하는데 있어서 다음과 같이 여러 가지 문제점들이 발생했다.

1) 일관성 유지가 어렵다.

여러 작업자들이 하는 작업이므로 일관성을 유지하기가 어려웠다.

2) 패턴 현상이 매우 다양하기 때문에 지침서만을 가지고는 다양한 정보를 표현하기 힘들었다.

3) 언어정보 구축 도구가 부족하다.

고정표현을 구축하기 위해서 도구를 따로 개발하지는 않았다. 다만, 본 논문에서는 한국어 형태소 해석기와 품사 태거를 이용하여 한국어 원시 문장의 형태소 분리를 쉽게 하도록 하였다. 지식베이스를 확장하고 일관성을 유지하기 위해서는 패턴 작업을 위한 도구가 개발되어야 한다.

4) 구축하고자 하는 고정표현과 한영 사전의 용례와의 차이가 크다.

한영 사전의 대역 용례는 의미전달이 목적이고 영어다운 표현을 사용하기 때문에 의역이 많다. 따라서 고정표현을 구축할 때 한국어와 영어에서 대응되는 단어를 찾을 수 없는 경우가 너무 자주 발생된다. 더구나 은유적 표현인 경우에는 패턴을 찾을 수가 없다.

KS: 그들은 사이가 가까워지기 시작했다
ES: A warm friendship sprang up between them.
KP: A=그들!이 사이!가 가까워지!기 시작하!다;
PAST,u
EP: A warm friendship spring\u up between
A=them

5) 한국어 명사가 영어에서 동사로 표현되는 경우

이와 같은 용례에서는 대응되는 명사가 없으면 메타 어휘를 사용하지 않는다.

KS: 목이 갈리도록 고함을 지르다
ES: shout oneself hoarse.
KP: A=그!가 목!이 갈리!도록 고함!을 지르!다\u
EP: A=He shout\u A:REFL=himself hoarse.

6) 한국어 동사가 영어 문장에서 명사로 표현되는 경우

이 경우는 한국어에서는 동사로 표현된 단어가 영어 문장에서는 명사나 그 밖의 다른 품사로도 표현되는 경우가 있다.

KS: 가르치면 보람이 있다
ES: Education pays.
KP: 가르치!면 보람!이 있!다
EP: Education pay\u

7) 한국어와 영어의 품사가 일치하지 않는 경우

아래의 용례에서 "중병"이라는 한국어 명사가 영어에서는 형용사로 사용되었다. 이 경우에도 한국어 명사에 대해서 메타 어휘를 사용하지 않는다.

KS: 그는 중병으로 매우 위태로운 지경에 있다
ES: He is very ill
KP: A=그가 중병으로 매우 위태롭ㄴ 지경에 있다
EP: A=He be very ill and A:POSS=his life hang by a thread

8) 한국어 예문에 나타나는 명사가 영어 문장에서는 나타나지 않는 경우

아래의 예문에서 한국어 명사 "말"에 대응되는 영어 단어를 찾을 수가 없다. 한국어 명사에 대응되는 영어 명사가 없기 때문에 메타 어휘를 사용하지 않는다.

KS: 그의 말은 아무래도 위태롭다
ES: He is not to be relied upon
KP: A=그의 말이 아무래도 위태롭다
EP: A=He be not to be relied upon

9) 한국어 명사가 여러 개의 영어 단어로 대응되는 경우

아래의 예문에서는 한국어 "국력"이 영어에서 두 개의 단어에 해당되므로 대응시킬 수 없다.

KS: 국력이 커지다
ES: a nation grows in power
KP: 국력이 커지다
EP: a nation grows in power

10) 한국어 명사가 문장으로 표현되는 경우

이 경우는 대부분 패턴으로 표현되지 않았다. 아래의 예문에서는 한국어 명사 "결석자"가 관계절을 이용해서 표현되고 있다.

KS: 결석자는 불합격으로 간주한다
ES: Those who absent themselves will be considered to have failed in the examination
KP: A=그를 불합격으로 간주하다
ES: A=He be considered to have failed in the examination

11) 완전 의역되어 일반적으로 사용되는 명사의 대응관계와 다를 경우

전체 문맥 상황에서는 뜻이 전달되지만 완전하게 의역되어 명사들의 대응이 불명확하게 된다. 아래의 예문에서 "가게"에 해당되는 단어가 "shop"인데 "거들다"라고 하는 동사 때문에 "일"이라는 단어가 추가되었다.

KS: 가게 일을 거들다
ES: help in the shop
KP: A=가게_일을 거들다
EP: help in A=the_shop

4. 결 론

본 논문에서는 한영기계번역 시스템에서 사용하기 위한 고정표현 지식 베이스를 구축 과정을 기술하고, 고정표현 지식을 기술하면서 발생하였던 여러 가지 문제점을 예와 같이 기술하였다.

빈도수에 의해서 선정된 5,000 개의 용언에 대하여 시사영어사의 엘리트 한영대사전으로부터 용례를 수집하여 고정표현 지식의 기술 방법에 따라 58,000여 개의 고정표현 지식을 구축하였다.

사전에 있는 용례들은 의역이나 은유적인 표현이 많고 한국어와 영어의 언어적 차이 때문에 대응되는 단어가 없거나 품사가 달라서 고정표현 지식을 구축하기가 매우 어려웠다.

앞으로 고정 표현 지식 베이스를 확장하고 일관성을 유지하기 위해서는 고정 표현 지식 구축을 위한 전용의 작업 도구를 개발하여야 한다.

참고 문헌

[1] 박상규 (1999) 국내의 기계번역 동향 및 자동번역 방법론 분석, 자연언어처리 튜토리얼, 고려대, pp.89-125
[2] 서병락 (1996) 한영기계번역을 위한 번역 패턴 기반의 영어 문장 생성, 서울대학교, 컴퓨터공학과 박사학위 논문.
[3] 이하규 (1994) 숙어 분산 특성을 이용한 한영 번역 숙어의 인식, 서울대학교, 컴퓨터 공학과, 박사학위 논문.
[4] 임철수, 김길창 등 (1997) 어휘화된 규칙에 기반한 영한기계번역시스템, '97 한국정보과학회 가을 학술대회 발표 논문집(II), pp.161-164
[5] 최명석 (1998) 한국어 부분 구문 분석: 코퍼스로부터의 규칙 자동 추출, 한국과학기술원, 전산학과, 석사학위 논문.
[6] 김재훈 외 5인 (1999) KTAG99: 새로운 환경에 쉽게 적응하는 한국어 품사 태깅 시스템, 제1회 형태소분석기 및 품사 태거 평가 워크숍, pp.99-105
[7] 안동언, 김재훈 (1999) 한영 기계번역을 위한 고정표현 지식 개발, 최종보고서, 한국전자통신연구원