한영 기계번역을 위한 고정표현 지식의 기술 방법

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

The Representation Method of Fixed Expression Knowledge for Korean-to-English Machine Translation

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

요 약

규칙기반 기계번역 시스템의 문제점을 보완하고자 제시된 예제기반 기계번역 시스템은 대량의 고품질 대역 코퍼스가 필요하다. 그리고, 빠른 N-best 예제 검색, 유사 예제 계산, 번역결과의 평가 등이 중요한 문제들이다. 또한, 무엇보다도 기본적인 것은 대역 예문들을 표현하고 기술하는 방법이다.

본 논문에서는 자연어 대역 예문들을 수집하여 기계번역 시스템에서 사용하는 고정 표현 지식을 기술하는 방법에 대해서 논의한다. 대역 패턴의 기술 방법을 CFG 형태로 정의하고 실제 용례를 통하여 기술 방법을 설명한다.

1. 서 론

규칙기반 기계번역 시스템의 문제점을 보완하고자 제시된 예제기반 기계번역 시스템은 대량의 고품질 대역 코퍼스가 필요하다. 그리고, 빠른 N-best 예제 검색, 유사 예제 계산, 번역결과의 평가 등이 중요한 문제들이다. 또한, 무엇보다도 기본적인 것은 대역 예문들을 표현하고 기술하는 방법이다[1].

예제기반 기계번역시스템을 위한 대량의 예문을 구축하기 위해서는 대역 코퍼스로부터 다양한 예문들을 수집하여야 한다. 단순히 수집하여 나열해 놓은 예문들을 직접 기계번역시스템에서 사용할 수는 없으며 기계번역에서 사용할 수 있는 형태로 기술되어야 한다.

기계번역에서 번역할 때 중심이 되는 단어는 용언이다. 또한 용언들은 필수항인 명사구들을 수반하고 있다. 이러한 패턴들은 용언에 따라 고정적이므로 고정표현이라고 정의한다.

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

본 논문에서는 이러한 고정표현의 기술 방법을 CFG 형태로 정의하고 실제 용례를 통하여 기술 방법을 설명한다. 코퍼스로부터 이렇게 기술된 한영 대역 패턴은 한영기계번역에 사용된다.

2. 연구 현황

예제기반 기계번역시스템이 새롭게 대두된 것은 대량의 단일언어 코퍼스 및 대역 코퍼스의 이용이 가능해졌기 때문이다[1][2].

정렬시스템은 대역 코퍼스에서 서로 대응되는 대역 어휘를 자동으로 추출하는 시스템이다[2]. 예제기반 한영 기계번역에 한영 정렬시스템을 이용할 수도 있지만 이 시스템은 예문보다는 어휘에 중점을 두기 때문에 예제기반 번역시스템에 직접 이용하기가 어렵다.

Pennsylvania 대학의 Joshi 팀은 TAG(Tree Adjoining Grammar)를 기반으로 한국어의 TAG를 영어의 TAG로 대응함으로써 기본적인 기계번역 시스템을 구성하고 있다[3]. TAG에는 많은 어휘적 정보를 표현할 수 있지만 TAG에 표현되는 어휘적 정보는 구문적인 정보 중심으로 하고 있기 때문에 본 논문에서 구축하고자 하는 어휘

<대역 패턴> → <한국어 토큰리스트> > <영어토큰리스트>
<한국어토큰리스트> →<실질형태소:형식형태소:분석 정보> <한국어토큰리스트> |
<메타 어휘=실질형태소:형식형태소:분석 정보> <한국어토큰리스트> |
<메타 어휘!형식형태소:분석 정보> <한국어토큰리스트> |
NIL
<영어토큰리스트> →<일반 어휘> <영어토큰리스트> |
<메타 어휘:활용 정보=일반 어휘:생성 정보> <영어토큰리스트> |
<메타 어휘:활용 정보:생성 정보> <영어토큰리스트> |
NIL
<메타 어휘> → A | B | ... , X | Y |···
(체언은 A,B 순으로, 용언은 X,Y순으로 기술한다)
<분석 정보> →<시제 정보> | <양상 정보> | <문장 정보> | <화법 정보> | <품사 정보>
<생성 정보> →<시제 정보> | <양상 정보> | <문장 정보> | <화법 정보> | <품사 정보>

<그림 1> 고정표현 지식의 기술 방법

정보가 첨가된 표현 지식을 기술하는 수단으로는 부족하다.

기존의 연구 중에서 고정표현과 유사한 것으로 서울대에서 연구한 숙어 표현[3][4]과 KAIST에서 연구한 chunk[5][6]가 있다. 본 논문의 고정표현은 이러한 연구들을 반영하였으며 숙어뿐만 아니라 언어적 성질도 함께 내포하여 해석과 변환에서 모호성을 최소화하도록 하였다.

3. 고정 표현 지식 베이스 구축

3.1 구축 대상 용언 선정

대상 용언은 한국어 뉴스 코퍼스에서 뽑은 가장 높은 빈도를 가지는 5000개를 선정하였다. 예문은 시사 엘리트 영한사전에서 대상 용언에 나타난 문장을 사용하였다.

3.2 대역 패턴의 종류

대역 패턴은 크게 용언형과 문장형 대역 패턴으로, 구분한다. 용언형이란 하나의 한국어 용언으로 이루어진 패턴을 말하며, 문장형이란 둘 이상의 한국어 용언으로 이루어진 패턴을 말한다.

용언형 대역 패턴

- *A:의 B=손:을 꼭 잡:다*
  > squeeze A:POSS B=hand

문장형 대역 패턴

- *아무리 하:여도 A:를X=잡아두:ㄹ 수:가 없:다*
  > Nothing can keep A from X=leaving

3.3 대역 패턴 기술 형식

기계번역 시스템에서 대역 패턴이 해석과 변환에서 직접 사용되기 때문에 대역 패턴의 기술은 고정 표현 지식을 구축하는데 가장 중요한 부분이다. <그림 1>은 고정표현 지식의 기술 방법을 CFG 형태로 정의하였다. 이 기술 방법을 기초로 다음의 규칙을 이용하여 대역 패턴을 기술하였다.

3.3.1 대역패턴의 구축은 가능한 사전에 나온 용례에 기초로 작성한다. 그러나, 용례에는 나타나지 않은 어휘를 복원하여 기술할 수도 있다.

용 례 : 길을 잃다 lose oneself
패 턴 : A:가 길:을 잃:다 > A lose A:REFL
용 례 : 호감이 가다 be favorably disposed toward
패 턴 : A:에게 호감:이 가:다
> be favorably disposed toward A

3.3.2 한국어 어휘에 대해, 그 어휘가 다른 어휘로 대치되더라도 동일한 영어 패턴으로 번역될 경우, 메타 어휘로 기술한다.

- *A=서울:에 가:다 > go to A=Seoul*

3.3.3 한국어 패턴에서 메타 어휘로 기술하여도 영어 패턴에서 대응되는 메타 어휘가 나타나지 않거나 적절하지 않은 경우에 한하여 일반 어휘로 기술한다.

- *A:가 혼자 가:다:CAN*
  > A can find A:POSS way
- *납득:이 가:다 > be convinced*

3.3.4 동사의 바로 뒤에 위치하는 메타어휘의 격은 목적격, 앞에 위치하는 격은 주격임을 가정하며, 이 경우 활용정보를 명시하지 않는다. 이 외의 경우는 체언에 활용정보를 명시하여야 한다.

[표 1] 체언의 활용정보

체언의 격	활용정보
주 격	SUBJ
간 접 목 적 격	IOBJ
직 접 목 적 격	DOBJ
소 유 격	POSS
재 귀 대 명 사	REFL

- *A:가 B=연필:을 잡다*
  > A take B=a_ pencil in A:POSS hand
- *A:가 길:을 잃:다 > A lose A:REFL*

3.3.4 용례에 나타난 한국어 용언이 시제나 양상정보를 가지고 있다면 다음과 같은 활용정보를 기술한다.

[표 2] 용언의 자질 정보

값	정 보
시 제	PRES, PAST, FUTU, PPAST
양 상	PROG, PERF
문 장	DECL, QUES, EXCL, INPR, SUGG
화 법	CAN, WILL, MUST, NEED, WISH, USEDTO, SEEM etc...

3.3.5 용언의 활용정보를 아래와 같이 기술한다.

[표 3] 용언의 활용 정보

용언의 생성	활용정보
용 언 의 원 형	BSE
to - 부 정 사 형	INF
동 명 사 형	GER
현 재 분 사 형	PSP
과 거 분 사 형	PRP
비 교 급	COMPRA
최 상 급	SUPRA

- *A:를 기대하다 > look forward to A (O)*
  > look forward to A:GER (O)

3.3.6 활용이 일어나게 될 용언을 파악하기 위한 품사정보를 기술한다. 만일, 본동사와 조동사가 있을 경우는 조동사에 품사 정보가 붙게 된다.

- *A=소풍:을 가다 > go on A=a_picnic*
- *A=노령:으로 B=이마:에 주름살:이 가다*
  > A=Age make:v lines on B=forehead

3.4 고정 표현 지식 구축

선정된 용언을 기반으로 사전에 추출한 예문들에 대하여 고정표현 지식을 구축한다. 모든 예문은 용언을 중심으로 패턴을 기술하며, 원문과 대역 패턴을 <그림 2>와 같이 모두 기술한다.

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

<그림 2> 패턴 구축 예

4. 결 론

본 논문에서는 예제기반 한영기계번역 시스템에서 사용하기 위한 고정표현 지식의 기술 방법을 CFG 형태로 정의하였다. 또한 이 기술 방법을 기초로 대역 패턴을 구축하는 규칙을 설명하였다.

출현빈도에 의해서 선정된 5,000개의 용언에 대하여 시사영어사의 엘리트 한영대사전으로부터 한영 용례를 수집하여 58,000여 개의 고정표현 지식을 구축하였다.

하지만 양질의 번역 품질을 얻기 위해서는 지금보다 더 많은 고정표현 지식베이스가 요구되며 일관성있게 확장되어야 한다. 또한 지식구축에 많은 시간이 소요되므로 지식 구축을 위한 전용 작업 도구가 필요하다.

참고 문헌

[1] 박상규 (1999) 국내의 기계번역 동향 및 자동번역 방법론 분석, 자연언어처리 튜토리얼, 고려대, pp.89-125

[1] 한국과학기술원 (1997) 통합국어정보베이스, 국어정보처리기술개발 제3차년도 최종보고서, 과학기술처

[3] Egedi, D., Palmer, M., Park, H. S., and Joshi, A. K. (1994), Korean to English translation using synchronous TAGS, Proceedings of the Conference of the Association for Machine Translation in the Americas, Columbia, Marylang, pp. 48-55.

[3] 김나리 (1997) 패턴정보를 이용한 한국어 구문분석, 서울대학교, 컴퓨터공학과 박사학위 논문.

[4] 서병락 (1996) 한영기계번역을 위한 번역 패턴 기반의 영어 문장 생성, 서울대학교, 컴퓨터공학과 박사학위 논문.

[5] 이현아, 이공주, 김길창 (1997) 자연스러운 번역을 위한 두 단계 영-한 변환시스템, '97 한국정보과학회 가을 학술대회 발표 논문집(II), pp.181-184

[6] 임철수, 김길창 등 (1997) 어휘화된 규칙에 기반한 영한기계번역시스템, '97 한국정보과학회 가을 학술대회 발표 논문집(II), pp.161-164