제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

대화체 자동번역 시스템에서 대화상대 맞춤 존대표현 생성에 관한 연구

최승권*, 김영길
*한국전자통신연구원 언어처리연구팀
e-mail:{choisk, kimyk}@etri.re.kr

A Study on Generation of Polite Expressions for Dialogue Participants in Machine Translation System[1)]

Sung-Kwon Choi*, Young-Gil Kim*
*Natural Language Processing Team, ETRI

요 약

현재의 자동번역 방식의 문제점은 대화 상대에 상관없이 항상 일정한 존대 표현을 생성하여 자동번역 결과를 부자연스럽게 만들고 앞뒤 대화 문맥을 혼란하게 만든다는 것이다. 예를 들어 대화 상대가 달라지면 동일한 원문에 대해서도 자동번역 결과는 다른 존대 표현을 생성해야 하나, 현재의 자동번역 시스템은 항상 하나의 일관된 존대 표현을 생성한다. 이 이유는 자동 번역 시스템에서 사용하는 번역 지식 또는 데이터가 고정되어 있어 유동적으로 변하지 않기 때문이다.
본 논문에서는 이러한 기존 자동번역의 문제점을 해결하기 위하여, 소셜 네트워크(social network)에서 제공하는 디지털 인맥 정보와 같은 비언어적 정보와 발화상의 표현과 같은 언어적 정보로부터 대화 자간의 존대 관계를 계산하여 자동번역 결과에 반영함으로써 언어 문화적 존대 차이를 자동으로 극복하는 대화 상대 맞춤형 존대표현 자동 번역 방법을 기술하는 데 그 목적이 있다.

1. 서론

한국전자통신연구원(이후 ETRI)에서는 언어 장벽을 해소하기 위해 다양한 분야의 자동번역 시스템을 개발하여 왔다. 개발한 자동번역 시스템은 필요로 하는 기관에 기술 이전되어 활발히 사용되고 있다. 특허 문서 자동번역 서비스[1], 과학 기술 논문 자동번역 서비스[2], 군사용 장비 매뉴얼 자동번역 서비스[3]등이 그런 예이다. 2010년도부터는 전자우편과 메신저를 대상으로 한 자동번역 시스템을 개발하기 시작하였으며 그 성과에 대해서는 [4]에서 기술한 바 있다.

현재 전 세계적으로 자동번역 시스템의 번역 방식은 크게 규칙기반 자동번역 방식과 통계기반 자동번역 방식으로 나뉘어 있다. 통계기반 방식이 규칙기반 방식에 비해 개발 시간을 절약할 수 있으며 특정 언어에 제약 없이 자동번역 시스템을 개발 할 수 있다는 장점이 있는 반면, 이종 어족 간의 번역 품질은 규칙기반 방식 보다 뒤지는 경향이 있다. 그 이유는, 첫째는 자료 희소성 때문이며, 두 번째는 언어적 차이 때문이다[5]. 이 때문에 영한 자동번역기는 통계기반 자동번역 방식보다는 패턴을 규칙으로 한 패턴기반 자동번역 방식을 채택하고 있다.

2010년부터 메신저 번역 시스템을 개발하면서 접하게 된 문제점은 대화 상대에 따라 존대 표현이 변하여야 하나 대화 상대에 상관없이 항상 일정하게 자동번역 결과가 생성되는 것이었다.

본 논문의 목표는 패턴기반 자동번역 시스템에서 대화 상대에 따라 올바른 존대 표현이 생성되는 방법에 대해 기술하는 것이다.

2. 대화 상대와 존대 등급의 매핑

2.1. 한국어와 영어의 존대 실현

한국어와 영어의 존대 실현은 크게 어휘적으로 문법적으로 나눌 수 있다. 그 예들을 보면 다음의 도표와 같다.

<표 1> 한국어와 영어의 존대 실현

언어	존대	예(보통<->존대)
한국어	어휘적	밥<->진지, 먹다<->잡수시다, 주다<->드리다, 묻다<->여쭙다, 그<->그분, 나<->저/제
	문법적	접사:~님, 조사:~께서, 선어말어미:~시, 어말어미: 해라체-해체-하게체-하오체-해요체-합쇼체, 단순의문형<->부정의문형, 명령형<->의문형
영어	어휘적	What's up?<->How are you?, Sit down<->Sit down, please
	문법적	Want to join us?<->Would you like to join us?

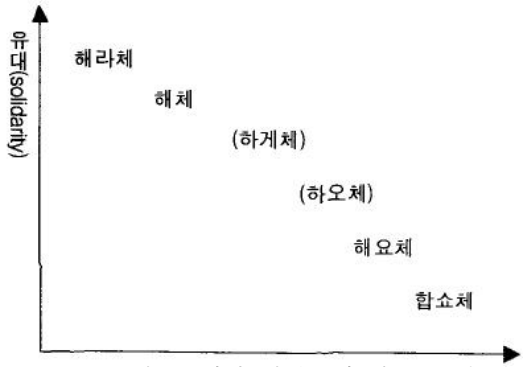
1) 본 논문은 지식경제부의 산업원천기술 개발사업 (2011-S-034-01)의 일환으로 개발된 결과임을 밝힙니다.

2.2. 존대에 영향을 끼치는 관계 요인들

존대에 영향을 끼치는 대화 상대들 간의 관계 요인은 크게 나이, 친밀도, 사회적 힘으로 나누어 볼 수 있다[6].

<표 2> 나이와 친밀도에 따른 존대 실현

대우 정도	구분	유소년기 (초등학생이하)	청년기 (10~20대)	장년기 (30~40대)	중년기 (50~60대)	노년기 (70대이후)
높임	손위	하세요체 해요체	하십시오체 하세오체 해요체	하십시오체 하세오체	하십시오체 하세오체	하십시오체
	동년배	해체	해요체 해체	하세오체 해요체 하게체 해체	하세오체 하시오체 하오체 하게체 하시게체 해체	하시오체 하오체 하시게체 하게체
	손아래	해체	해체	해체	하시게체 하게체 해체	하시게체 하게체 해체
안높임	손위	해체	해체	해요체	하세오체 하시오체 해요체	하시오체 하오체
	동년배	해체	해체 해라체	하게체 해체	하오체 하게체 하시게체 해체	하시게체 하게체
	손아래	해체	해라체	해체	하게체 해체	하시게체 해체



(그림 1) 힘과 친밀도에 따른 존대 실현

표 2는 각 연령대의 화자가 자신보다 손위, 동년배, 손아래 청자 중 어떠한 사람을 만나는가에 따라 그리고 친밀도에 따라 존대 표현이 달라짐을 보이는 것이다. 그림 1은 X축의 힘과 Y축의 친밀도에 따라 존대 표현이 바뀔 수 있음을 보여준다.

2.2. 존대 요인들과 존대 등급의 매핑

앞 절에서 언급한 한국어의 존대 실현을 고어체에 나타나는 존대 표현을 제외하고 단순화 시키면 “해체”,“해요체”,“합쇼체”로 만들 수 있다. 두 대화자의 대화상에 나타나는 존대표현 실마리를 토대로 한국어의 어휘적 존대와 문법적 존대에 의한 언어표현 실마리와 존대 등급과의 매핑을 만들면 다음과 같을 수 있다:

<표 3> 존대표현 실마리와 존대등급의 매핑

어휘적 존대	접사	조사	선어말어미	어말어미	존대등급
보통어휘	-	-	-	해체	0
보통어휘	-	-	-	해요체	1
존대어휘	~님	-	~시	해요체	2
존대어휘	~님	~께서	~시	합쇼체	3

또한 인간관계 존대 요인들인 {나이, 친밀도, 힘}의 조합과 존대 등급과의 매핑은 다음과 같다:

<표 4> 인간관계 존대 요인의 조합과 존대등급의 매핑

대화자1이 대화자2보다 나이가	대화자1과 대화자2의 친밀도가	대화자1이 대화자2보다 힘이	대화자1 존대등급	대화자2 존대등급
많다	높다	크다	2	0
많다	높다	작다	2	1
많다	높다	모른다	1	0
많다	낮다	크다	3	1
많다	낮다	작다	2	2
많다	낮다	모른다	3	1
많다	모른다	크다	3	0
많다	모른다	작다	3	3
많다	모른다	모른다	3	1
적다	높다	크다	1	2
적다	높다	작다	0	2
적다	높다	모른다	1	2
적다	낮다	크다	1	3
적다	낮다	작다	1	3
적다	낮다	모른다	1	3
적다	모른다	크다	3	3
적다	모른다	작다	2	3
적다	모른다	모른다	3	3
모른다	높다	크다	2	1
모른다	높다	작다	1	2
모른다	높다	모른다	2	2
모른다	낮다	크다	3	2
모른다	낮다	작다	2	3
모른다	낮다	모른다	3	3
모른다	모른다	크다	3	2
모른다	모른다	작다	2	3
모른다	모른다	모른다	3	3

이 밖에 존대 표현 실마리나 인간관계 존대 요인 보다는 직관적인 인간관계에 의한 존대 등급과의 매핑을 예로 들면 다음과 같다:

<표 5> 직관적인 인간관계와 존대등급의 매핑

관계의 예	존대등급
친구, 친한 후배, 친한 제자	0
친한 선배	1
친한 교수, 친한 상사	2
모르는 연장자, 친하지 않은 상사	3
기타	

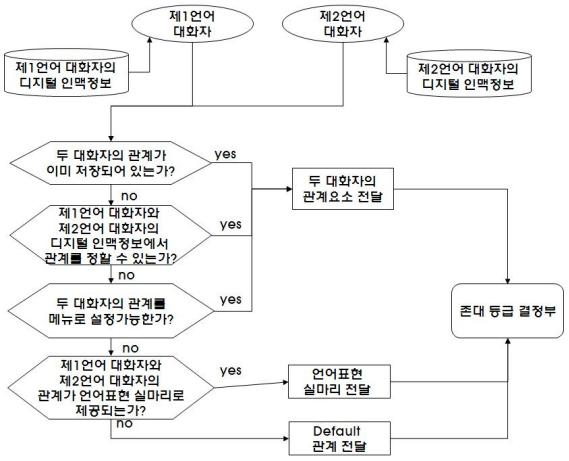
3. 두 대화자의 관계 인식 흐름도

대화 상대를 인식하는데 있어 상황인식 시스템에 근거한 방법[7]이 제시된 바 있다. 본 논문에서는 Facebook이나 Twitter와 같은 쇼셜 네트워크(social network)로부터 두 대화자의 공개된 프로필과 같은 디지털 인맥정보를 사용하여 두 대화자의 관계를 추정하는 방법을 제시하고자 한다. 공개된 개인 정보를 사용하는 이유는 개인 정보를 무단으로 사용하는 것은 정보통신법의 개인정보 사용에 저촉되기 때문이다.

두 대화자의 관계를 인식하는 방법은 두 대화자의 공개된 디지털 인맥 정보를 토대로 단계적으로 파악할 수 있다. 전체 흐름도는 그림 2와 같다.

‘나이’, ‘친밀도’, ‘힘’과 같은 두 대화자의 관계요소나

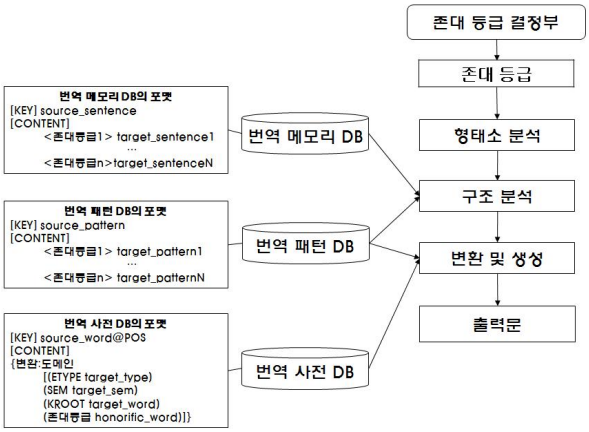
'호칭', '존대표현'과 같은 언어표현 실마리나 '격식'과 같은 디폴트 관계 정보를 받아서 2장에서 기술한 매핑에 의해 존대 등급을 만들 수 있다.



(그림 2) 두 대화자의 관계 인식 흐름도

4. 자동 번역에서의 존대 등급 반영 방법

존대 등급 결정부에 의해 존대 등급이 결정되면 이 존대 등급은 각 모듈별로 전달되면서 적절한 존대 표현이 선택되게 된다. 구조분석에서는 번역 메모리나 번역 패턴의 존대 표현이 선택되게 되며, 변환 및 생성에서는 번역 사전의 존대 대역어가 선택되게 된다. 전체적인 시스템 구성도는 다음과 같다.



(그림 3) 대화상대 맞춤 패턴기반 자동번역 시스템 구성도

시스템 구성도에 따른 자동번역기의 존대 처리 과정을 예를 들어 설명하면 다음과 같다:

(그림 4) 자동번역에서의 존대처리 과정의 예

그림 4에 따르면, 1단계에서는 어휘번역 패턴 및 번역메모리 DB에서 한국어 형태소 분석기에 의해 한국어의 어미 제거 및 본동사에 핵심어 표시를 준비해야 한다. 즉 "I can tell you the web site <-> 웹사이트는 말해 줄 수 있다."라는 번역메모리 예문에서 한국어 대역문의 핵심 동사인 "말해 줄 수 있"에 핵심어 표시 !를 표시함으로써 존대 등급이 전달되게 된다. 2단계에서는 두 대화자의 관계를 반영하기 위해 그림 2의 관계 인식 흐름도에 따라 존대 등급 계산 테이블을 이용한다. 각 등급은 앞서 2장에서 언급한 등급에 의해 언어적 실현을 만드는 것이다. 3단계에서는 존대 등급에 따른 두 대화자의 존대 어미를 자동으로 생성하게 된다. 두 대화자의 관계에 따라 적절한 존대 어미가 생성된 결과는 다음과 같을 것이다

[모르는 사이]
- 존대등급: 3
- 존대실현: 존대어휘, 합쇼체

I can tell you the web site <-> 웹사이트는 말해 줄 수 있습니다.

[친구 사이]
- 존대등급: 0
- 존대실현: 보통어휘, 해체

I can tell you the web site <-> 웹사이트는 말해 줄 수 있어.

5. 결론

현재의 자동번역 시스템들은 사용하는 번역지식 또는 데이터가 고정되어 있어 유동적으로 변하지 못하기 때문에 대화 상대가 변화더라도 항상 일관된 존대 표현을 생성하는 문제점을 가지고 있었다. 본 논문에서는 이러한 기존의 문제점을 해결하기 위해, 소셜네트워크에서 제공하는 디지털 인맥 정보와 같은 비언어적 정보와 발화상의 표현과 같은 언어적 정보로부터 대화 상대에 따른 존대 관계를 계산하여 자동 번역 결과에 반영함으로써 언어 문화적 존대 차이를 자동으로 극복하는 대화 상대 맞춤형 자동 번역 생성 방법을 기술하였다. 따라서 본 논문에 따른 자동 번역 시스템에서는 대화 상대에 따라 자동 번역 결과가 고정된 표현이 아닌 대화 상대에 맞는 존대 표현이 생성된다.

참고문헌

[1] Sung-Kwon Choi, Oh-Woog Kwon, Ki-Young Lee, Yoon-Hyung Roh, and Young-Gil Kim. "Customizing an English-Korean Machine Translation System for Patent Translation", In Proceedings of the 21st Pacific

Asia Conference on Language, Information and Computation PACLIC 21), 2007, pp.105-114.

[2] Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, Oh-Woog Kwon, and Young-Gil Kim. "How to Overcome the Domain Barriers in Pattern-Based Machine Translation System", In Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation(PACLIC 22), 2008, pp.161-168.

[3] Oh-Woog Kwon, Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, and Young-Gil Kim. "Customizing an English-Korean Machine Translation System for Patent/Technical Documents Translation", In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation(PACLIC 23), 2009, pp.718-725.

[4] 최승권, 이기영, 노윤형, 권오욱, 김영길. "문어체에서 대화체 문장 패턴기반 영한 번역기로의 특화", 한글 및 한국어 정보처리, 2010, pp.136-140.

[5] 최승권, 김영길. "영한 번역기의 상용화를 위한 도메인 특화 방법의 진화", 한글 및 한국어 정보처리, 2011, 게재 예정.

[6]권오병, 최석재, 박태환. “존대등분 계산법과 사례기반 추론을 활용한 상황 인식형 모바일 인터페이스 시스템”, 한국 지능정보 시스템 학회 논문지, 제13권 제3호. 2007, pp. 141-160.

[7]박태환. “상황인식시스템을기반으로한예의바른에이전트와 사용자인터페이스”, 한국경영정보학회 춘계학술대회, 2009, pp.891-896.