영한 대화체 자동번역을 위한 특화 방안

이기영,○ 노윤형, 권오욱, 최승권, 김영길
한국전자통신연구원 언어처리연구팀
{leeky, yhroh, ohwoog, choisk, kimyk}@etri.re.kr

Customization for English-Korean Spoken Language Machine Translation

Ki-Young Lee○, Yoon-Hyung Roh, Oh-Woog Kwon, Sung-Kwon Choi, Young-Gil Kim
Natural Language Processing Team, Electronics and Telecommunications Research Institute

요 약

현재, 자동번역의 도메인은 응용 프로그램의 요구에 따라, 점차 문어체에서 대화체(spoken language)로 옮겨가고 있는 추세이다. 본 논문은 대화체가 지니는 특성을 자동번역 시스템을 구성하는 각 모듈별 및 지식 관점에서 분석하였다. 특성 분석을 기반으로 하여, 본 논문에서는 여행 영역을 대상으로 하는 대화체 자동번역시스템의 특화를 수행하였다. 대화체 자동번역을 위한 새로운 지식으로 구조화 번역메모리(Translation Memory)가 도입되었으며, 시스템을 구성하는 각 모듈별로 대화체 특화가 이루어졌다. 또한 기존의 문어체용 기구축 패턴 등이 정비되었으며, 고빈도 대화체 표현에 대한 신규 패턴이 도입되었다. 제안하는 방법의 검증을 위해 수동평가를 수행하였으며, 그 결과, 영한 대화체 자동번역에 있어서 번역률 향상이 있었다.

주제어: 대화체, 자동번역, 대화체 자동번역

1. 서 론

자동번역 기술은 지금까지 규칙 기반의 자동번역 방법론을 필두로 하여, 예제 기반, 패턴 기반 및 통계 기반의 방법론까지 다양한 방식으로 개발되어 왔으며, 자동번역이 대상으로 하는 문서의 도메인도 다양한 분야로 점차 확대되고 있다. 이와 함께, 현재는 네트워크 기술의 발달과 함께 자동번역을 필요로 하는 새로운 분야가 생겨나고 있는 추세이다. 이러한 대표적인 분야로는 자동통역, 온라인 채팅 등이 대표적인 예이다.

자동통역이나 온라인 채팅 등의 응용분야에서 사용되는 문장은 대화체의 특성을 지닌다. 일반적으로 대화체는 문어체와는 달리 비문법적인 요소들을 많이 포함하며, 문어체와는 다른 특성을 지닌다. 따라서 기존의 문어체를 대상으로 하는 자동번역시스템을 대화체에 바로 적용할 경우, 번역 성능은 크게 떨어진다.

본 논문에서는 현재 자동번역 추세에 맞게 대화체가 지니는 특성을 자동번역시스템을 구성하는 각 모듈의 관점에서 파악하고 이를 분석하여 대화체 자동번역시스템의 지식 및 모듈별 특화에 대한 내용을 기술한다.

2. 관련 연구

국내의 경우, 한일/일한 자동번역은 문어체와 대화체 간의 기술적 차이가 거의 없어서, 다른 언어쌍과 달리 쉽게 대화체 번역 기술을 적용한 번역 서비스가 국내 포털 사이트에서 제공되고 있다. 하지만, 한국어를 중심으로 영어와 같은 구조적 차이가 큰 언어 간의 대화체 자동번역 기술은 아직 연구가 진행 중에 있다.

국외의 경우, 대화체 자동번역을 위한 연구는 국내에 비해 상대적으로 더 오랫동안 진행되었으며, 통계 기반의 자동번역 방법론을 기반으로 한 대화체 자동번역을 위한 연구가 활발하게 이루어지고 있다. 구글(Google)의 경우, 자체 개발한 통계 기반 자동번역 엔진인 Google Translate를 구글 Talk 및 구글 e-mail에 적용하여 한국어를 포함한 24개 언어쌍에 대해 자동번역 서비스를 제공하고 있다. [1]은 기존의 통계기반 방식의 대화체 자동번역 기술을 PDA와 같은 모바일 기기에 적용하는 기술을 소개한다. [2]는 자동번역 기술을 자동통역에 적용하여 음성인식 및 음성합성과 결합된 자동번역 기술을 보인다. 국외의 경우, 대화체 자동번역을 위한 다양한 방법론이 보고되고 있으며, 이러한 기술을 모바일 기기 등에서 구현하는 다양한 응용 기술들이 개발되고 있다

3. 영어 대화체 특성 분석

영어 대화체의 특성 분석을 위해 사용한 코퍼스는 인터넷에서 수집한 여행 영역의 대화로 구성된 1천 문장을 사용하였다. 본 장에서는 영어 대화체가 갖는 언어적 특성을 시스템의 관점에서 다루며, 이렇게 분석된 대화체 특성은 자동번역시스템의 도메인 특화 및 튜닝을 위해 사용되었다.

3.1 번역 지식 관점

일반적으로, 대표적인 대화체의 특성은 문맥에 의한 생략 현상

과 대화 분야에 전형적인 대화체 표현일 것이다. 다음의 예문들은 대화체의 생략 현상과 대화 분야에 전형적인 대화체 표현들과 관련된 자동 번역에서의 어려움을 잘 나타낸다.

(대화1) 전화 대화

A: Hello, could I speak to Mr. John Jarrett, please? (여보세요, 존 제러트씨 좀 바꿔 주실래요?)

B: This is Mr. Jarrett speaking. (전데요)

(대화2) 호텔 예약

A: Do you have any rooms avaliable from October 9th through the 10th? (10월 9일부터 10일까지 사용할 방 있어요?)

B: Yes, we do. How many are in your party? (네, 있습니다. 몇 분이신데요?)

(대화3) 모닝콜

A: Will you give me a wakeup call at 6 AM? (내일 아침 여섯시에 깨워 주시겠어요?)

B: Certainly, sir. I'll leave a note with the operator. (그러세요, 손님. 교환원에게 연락해 두겠습니다.)

자동 번역 관점에서 볼 때 위의 전형적인 영어 대화 표현들의 한국어 번역은 쉽지 않다. (대화1)의 “This is Mr. Jarrett speaking"에 대응되는 전형적인 한국어 번역은 ”전데요“인데 현재의 영한 자동 번역 시스템에서는 이러한 번역을 생성할 수 없다. 또한 (대화2)에서 예약 일정에 대한 질문에 대해 ”Yes, we do."라는 답변은 “Yes, we have some rooms."의 생략 현상으로 ”Yes, we do."의 올바른 번역 표현인 “네, 있습니다.”를 현재의 영한 자동 번역 시스템에서는 생성할 수가 없다. (대화3)에서의 “Certainly"라는 부사에 대한 대역어가 문어체에서는 대부분 ”확실히“라는 한국어 대역어로 사용되기 때문에 문어체의 대역어를 대화체에서 그대로 사용하면 어색한 번역이 된다.

3.2 형태소 분석 및 품사태깅(Tagging) 관점

영어 형태소 분석은 크게 단어(형태소, 토큰) 분리와 분리된 단어에 대한 품사태깅으로 나눌 수 있다.

한국어와 달리, 영어에서는 단어를 분리 및 분석하는 과정이 복잡하지 않고 단순하다. 하지만, 자동번역을 위한 형태소 분석에서는 단어 분리 과정에서 구조분석과 한국어 생성 과정에 도움을 주는 호칭 표현, 날짜/시간/수/화폐 표현 청킹(Chunking) 등에 대하여 미리 처리하기를 요구한다. 이러한 전처리는 처리하고자 하는 도메인에 따라 다양하게 나타나며 처리 과정이 다를 수 있다. 단어 청킹 관점에서 문어체와 대화체의 차이점을 보면, 대화체는 다양한 날짜/시간/수/화폐 표현을 포함하며, 문어체의 구체적인 표현과는 달리 일부 요소가 생략되어 표현되기도 한다.

(예문1) Do you have any rooms avaliable from October 9th through the 10th?

(예문1)과 같이 정상적인 날짜 표현이 들어온 경우에는 단어 분리에서 “from October 9th through the 10th”를 보아, "October 9th"를 날짜 표현으로 인식하고 뒤의 “the 10th"에 대해서는 월 표현인 "October"가 생략된 것을 앞의 단어들을 통해 파악함으로써 ”October 10th"로 변경할 수 있다. 이렇게 처리함으로써, 추후 구조분석과 한국어 생성에서 "from 날짜 through 날짜“ 표현에 대한 구조분석과 생성이 정확하게 파악이 가능하다.

(예문2) Do you have any rooms avaliable from 9th through 10th?

하지만, 대화체에서는 년도나 월에 대한 표현을 생략하고 (예문2)와 같이 표현하는 경우가 빈번하다. 이러한 경우에는 “9th"와 ”10th"가 날짜인가 서수로 사용된 것인가가 불명확하여 형태소 분석 단계에서 처리하기가 곤란하다. 이와 같은 표현의 부분 생략은 날짜 이외에 시간표현과 화폐 표현에서 자주 발생한다.

수 표현의 경우에는 문어체보다 대화체에서 더욱 다양하게 사용한다. 분수나 수학식 등에 대하여 문어체에서는 숫자와 기호로만 표현한 것을 영어식으로 읽는 표현으로 나타나기 때문이다.

일반적으로, 형태소 품사태깅 관점에서 문어체와 대화체의 차이점은 분명히 존재한다. 우선 대화체의 문장 길이가 매우 짧고, 의문형/명령형/청유형 문장 표현이 많이 사용된다. 그리고, 감탄사 및 부사의 문두 사용이 빈번하며 생략이 빈번하다. 또한, 대명사와 조동사의 출현 빈도가 문어체와 비교하여 매우 크다. 그러므로, 품사태깅에서 사용하는 n-gram의 확률이 달라진다. 대화체에서 나타나는 n-gram이 문어체에서 아예 출현하지 않는 것이 아니라, 대부분이 문어체의 n-gram으로 출현하지만 그 정도가 다르다. 이러한 n-gram의 차이점은 명령형/청유형 문장과 주어 생략 문장에 대한 품사태깅의 오류를 일으킬 수 있다. 문어체에서는 문두에서 항상 체언류로 시작하는 경우가 많으나 대화체에서 용언류로 시작하는 경우도 많기 때문이다.

3.3 구조분석 관점

영어 대화체 문장의 구문 특성을 요약하면 표 1과 같다. 구조분석의 관점에서 대화체 문장을 분석하는데 특히 어려운 점은 문장 성분의 생략이다.

표 1. 대화체 문장의 구문적 특성

특성	문어체	대화체
문장 길이	길다	짧음
관용 표현	적음	상대적으로 많음
명령문/의문문	적음	많음
문장성분 생략	적음	많음
기타 성분의 사용	정규적임	감탄사, 호칭, 간투사를 포함함
고유명사의 사용	적음	많음

(예문3) Not with that!

(예문4) Not for carving you.

(예문3)과 (예문4)는 특히 문장 성분이 생략된 대화체 문장의 예를 보인다.

또한 대화체 문장은 문장 중간에 감탄사, 호칭, 간투사 등을 포함한다. 이러한 종류의 예로는 Oh dear!, darling, yeah man! well, you know, Jack, 등이 있다.

(예문5) Would you like breakfast?

(예문6) How much will it be?

(예문5)와 (예문6)은 규칙 기반의 구의 조합으로 자연스러운 대역어 생성이 어려운 문장의 예를 나타낸다. 마지막으로 대화체 문장은 고유명사를 많이 포함한다.

이와 같은 특성들을 종합적으로 고려할 때 대화체 자동번역을 위해서는 전통적인 방식의 구구조 규칙의 조합에 의한 파싱(Parsing)보다는 문형이나, 패턴에 의한 분석 및 변환이 필요하다. 또한 감탄사, 호칭, 간투사 등에 대한 처리, 고유명사, 시간명사 등의 인식 및 처리 등이 필요하다고 할 수 있다.

3.4 변환 및 생성 관점

영어 대화체 문장은 생략된 표현을 많이 포함한다. 이러한 생략 현상은 자동번역시스템의 구조분석에도 많은 영향을 주지만, 변환 과정에서도 매우 중요한 영향을 미치는 현상이다.

(예문7) For London

(예문7)은 주어 및 동사 등이 생략된 형태를 지닌다. 이러한 문장은 문어체 자동번역시스템의 경우, 단순히 전치사구의 형태로 번역을 수행하여 “런던을 위하여”라는 번역문을 생성한다. 하지만 대화체에서 (예문7)를 올바로 번역하기 위해서는 우선적으로 대화 문맥을 살펴보는 것이 중요하다. 즉, “For London"이라는 문장이 어떠한 문맥에서 생성되었는지를 파악하여 변환 및 생성 이전의 단계에서 생략된 요소를 복원하고 변환 및 생성 단계에 이러한 복원 정보가 반영되어 올바른 번역 결과를 생성해야 한다. 또한 이러한 정보를 사용하여 적당한 종결어미 등을 함께 생성해야 한다.

(예문8) Hope you get well soon.

(예문8)은 주어가 생략된 표현으로서 영어의 명령어 표현과 동일한 형태를 취한다. 따라서, 분석 모듈뿐만 아니라 변환 모듈에서도 이러한 문장이 취하는 양상 정보를 올바로 분석하고 변환해야 한다.

(예문9) How to get downtown.

(예문9)는 대화체에 특히 많이 등장하는 모호성이 많은 동사 ‘get'을 포함한다. 대화체 문장에서 등장하는 이러한 모호성 어휘는 올바른 대역어 선택 장치가 없다면 매우 부자연스러운 번역 결과를 생성한다.

(예문10) I didn't sleep well.

(예문10)은 기존의 문어체 자동번역시스템을 사용하면 “나는 잠을 잘 자지 않았다.”라는 문장을 생성한다. 문장의 각 어휘가 그대로 번역되었지만, 실제로 대화 관점에서 볼 때는 어색한 표현으로 번역되었다. (예문10)의 자연스러운 번역 결과는 “잠을 설쳤어요” 이다. 변환 및 생성 모듈의 관점에서 이러한 문장은 번역 메모리 또는 패턴에 의해 자연스러운 번역 결과를 얻을 수 있다.

즉, 변환 및 생성 모듈의 관점에서 대화체의 특성은 주로 변환 및 생성 이전 모듈인 형태소 분석 및 구조분석 모듈과 큰 연관성을 지니고 있으며, 이러한 특성은 변환을 위한 번역지식인 패턴이나 사전 등에 반영되어야 하며, 변환 및 생성 모듈에서도 대화체 특화를 통한 성능 개선이 필요하다.

4. 영한 대화체 자동번역시스템

4.1 영한 대화체 자동번역시스템 구조

그림 1은 영한 대화체 자동번역시스템의 구조를 나타낸다. 영한 대화체 자동번역시스템은 자동번역을 위한 주된 지식으로 영한사전과 영한 어휘패턴을 사용한다. 입력된 영어 문장은 형태소 분석 및 품사태깅 과정을 통해 숫자, 날짜 및 고유명사가 인식되고, 입력 문장을 구성하는 각 어휘의 원형이 복원된다. 구조분석 과정에서는 규칙과 패턴 기반의 구조분석이 수행되며, 구조분석을 위한 지식으로는 규칙과 패턴 등이 사용된다. 그리고 변환 및 생성 과정에서는 구조변환과 어휘변환을 수행하고 최종적으로 한국어의 문법에 맞는 한국어 번역문을 생성한다.

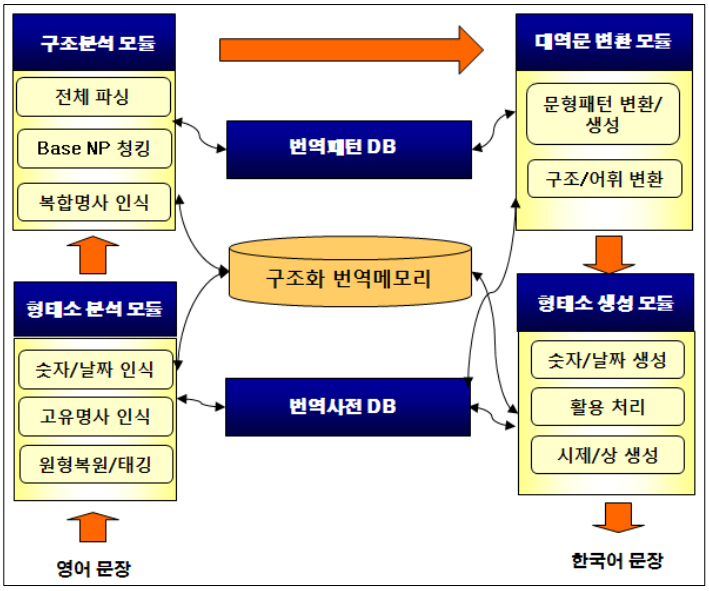

그림 1. 영한 대화체 자동번역시스템 구조

4.2 번역메모리 및 지식

4.2.1 구조화된 번역메모리

번역메모리란 원문의 문장과 그것의 번역된 문장을 하나의 쌍으로 하여 데이터베이스화한 것을 말한다. 이 번역메모리를 사용하는 목적은 번역가가 이전에 번역한 번역문을 재활용하려는 것이다. 그러나 번역메모리는 문자열(string)로만 기술되기 때문에 커버리지가 낮은 것이 단점이다. 구조화된 번역메모리는 이러한 기존의 번역메모리의 커버리지를 획기적으로 개선하기 위해 기존의 번역메모리에 언어학적 청킹(chunking) 및 정렬(alignment)을 도입한 새로운 번역메모리를 말한다 [3]. 구조화된 번역메모리를 반자동으로 구축하는 절차는 다음과 같다:

표 2. 구조화된 번역메모리의 반자동 구축 절차

단계	구축 방법	정렬
1. Sorting & Normalization	전처리	
2. Expansion	문두 부사 상당 어구 제거 및 확장	문두 부사 병렬 패턴 기반
3. Substitution	고유 명사 청킹 및 치환	고유명사 병렬 목록 기반
	숫자 청킹 및 치환	숫자 병렬 패턴 기반
4. Chunking	기본 명사구 청킹 및 Base NP 치환	기본 명사구 병렬 패턴 및 번역 사전 기반
	숙어 청킹 및 나머지 부분 변수 치환	수동

3.1절에서 기술되었던 예들을 구조화된 번역메모리의 형태로 기술하면 다음과 같다:

(대화4) 전화 대화

could i speak to mr. PRN[1] please? <-> PRN:씨 좀 바꿔 주실래요?

this is mr. PRN speaking. <-> 전데요.

(대화5) 호텔 예약

do you have BNP[2] avaliable from NUM[3]1 through NUM2? <-> NUM1:부터 NUM2:까지 사용할 BNP 있어요?

yes, we do <-> 네, 있습니다.

how many are in your party? <-> 몇 분이신데요?

(대화6) wakeup call

will you give me a wakeup call at NUM am? <-> 내일 아침 NUM 시에 깨워 주시겠어요?

certainly, sir. <-> 그러세요, 손님.

I will leave a note with BNP <-> BNP에게 연락해 두겠습니다.)

위의 구조화된 번역메모리에 의해 고유 명사, 날짜, 전화 번호, 기본 명사구, 숙어 표현들에 대해 커버리지를 높일 수 있다.

4.2.2 대역어 특화 작업

대화체에서 고빈도로 사용되는 영어 단어들에 대한 한국어 대역어 특화 작업이 이루어졌다. 고빈도 영어 단어들이 대화체 코퍼스로부터 자동 추출되었고 고빈도 영어 단어들 중에서 번역 사전에 대역어를 많이 가지고 있는 단어들에 대해 전문 번역가의 수작업으로 대역어의 특화가 이루어졌다 [4]. 예를 들어 certainly가 번역 사전에서 특화되기 전후의 모습은 그림 2와 같았다. 그림 2에서 "변환:WEB" 및 "변환:SPE"는 각각 도메인 정보를 나타내는 것으로서, WEB은 일반 웹 도메인 및 SPE는 대화체 도메인을 의미한다. 위에서 "ETYPE BS"는 문두 부사를 의미하며, “ETYPE BV" 동사 수식 부사를 의미한다.

특화 전	certainly@ADV {변환:WEB [(ETYPE BV)(SEM MANNER)(KROOT 확실히)]}
특화 후	certainly@ADV {변환:SPE [(ETYPE BS)(SEM MANNER)(KROOT 네)] [(ETYPE BV)(SEM MANNER)(KROOT 확실히)]} {변환:WEB

1) PRN: Proper Noun
2) BNP: Base NP
3) NUM: Numeric Expression

	[(ETYPE BV)(SEM MANNER)(KROOT 확실히)]}

그림 2. 번역 사전 특화 예

4.3 형태소 분석 및 품사태깅

기존의 문어체용 형태소 분석 및 품사태깅 모듈을 대화체에 적합하게 위해서는 크게 단어 분리 관점과 품사태깅 관점으로 구분하여 고려하여야 한다.

형태소 분석의 단어 분리 과정에서 대화체 처리를 강화하기 위하여, 다양한 수 표현에 대한 처리가 가능하게 하였다. 또한 날짜/시간/화폐 등의 표현에서 생략되거나 표현 순서가 바뀐 경우에 주위 실마리 단어와 품사 정보만으로 알 수 있으면 처리하도록 하였다.

문어체 형태소 품사태깅 모듈을 대화체 형태소 품사태깅 모듈로 특화하기 위해서는 당연히 품사태깅된 대화체 코퍼스에서 학습한 정보를 기존 정보에 대체하면 가장 효과적이다. 하지만, 품사태깅된 대화체 코퍼스를 구축하는 비용과 기간에 대한 문제점으로, 본 영한 자동번역 시스템에서는 몇 가지 특화를 통하여 기존 문어체 형태소 품사태깅 모듈을 대화체용으로 변경하였다.

문어체 형태소 품사태깅 모델을 대화체에 특화하기 위해서, 본 논문에서는 명령형과 청유형에 대한 품사태깅 오류를 수정하기 위한 후처리 규칙을 보강하였다. 예를 들면, please 뒤에는 거의 동사원형이 나오므로, please 뒤에 단수명사로 품사태깅된 단어를 동사원형이 가능하면 동사원형으로 변경하는 후처리 규칙을 추가하였다.

4.4 구조분석

주로 문어체 번역을 대상으로 했던 기존의 구조분석 모듈은 다음과 같은 구성으로 이루어졌다.

- 복합명사 인식
- 장문처리
- 청킹 패턴 인식
- 규칙, 패턴에 의한 파싱 수행

3.3절에서 언급한 바와 같이 대화체 특성을 고려한 구조분석을 위해서 다음과 같은 처리가 추가되었다.

- 규칙 및 패턴에 실시간 부사구 삽입 규칙 자동생성 및 추가: thanks for NP -> thanks ADVP* for NP를 추가함, 여기서 ADVP*는 선택적 부사구를 의미함, “Thanks again for your help”, “Thanks in advance for your help”와 같은 문장들이 매칭됨.
- 고유명사구 인식 및 청킹: Uncle_Fester (사람), Sun_Hotel (건물), Tiger_Woods (사람), Ala_Moana_Shopping_Center (건물)
- 문장 단위 번역메모리 적용: “thank you", "have a nice day", "what's up"
- 문형 패턴 적용: “i would like to make a reservation for NP”, “how long do it take to get from NP to NP by train”
- 주어 또는 주어+be동사, do+주어 생략을 위한 규칙 보완
- 청유문(let's), 명령문, 의문문을 위한 규칙 보강

위에서 문장 단위 번역메모리는 전통적으로 형태소 분석을 수행하기 전에 문자열 매칭하는 것과 달리 구조분석 과정에서 매칭을 수행한다. 구조 분석 중에 문장(S)로 인식된 범위에 대해 번역메모리 매칭을 수행함으로써, 문장에서 인용부호 안의 부분 문장에 대한 번역메모리 매칭을 수행하여 번역메모리의 적용률을 높일 수 있다.

이와 같은 분석 엔진의 보완에 더해서 구조 분석 성능을 결정하는 핵심은 어떻게 대화체 문장을 위한 문형이나 패턴들을 효과적으로 구축하느냐하는 것이다. 이를 위해 대화체 코퍼스에서 자동으로 가능성 있는 패턴 후보를 추출하여 사람의 노력을 최소화할 수 있도록 제시하여 주는 패턴 구축 도구가 필수적이다. 이를 위해 언어적, 통계적 방법을 적용한 패턴 추출 도구가 구현되었다.

4.5 변환 & 생성

4.5.1 패턴 구분 및 분리

변환 및 생성 모듈에서의 대화체를 위한 특화는 주로 번역지식의 특화와 함께 진행되어야 한다. 우선, 제안하는 영한 자동번역시스템은 패턴을 기반으로 하기 때문에, 기존의 문어체 번역을 위한 패턴과 대화체를 번역하기 위한 패턴을 분리하고 구분하였다. 기존의 문어체를 위한 패턴과 새로운 대화체를 위한 패턴이 혼재해 있을 경우, 서로 충돌을 일으키거나 혼용되어 오히려 번역 성능을 저하시키는 경우가 많기 때문이다. 따라서 특정 도메인에 특화된 패턴이 다른 도메인에 속하는 문장을 번역하는데 사용되지 않도록 함으로써, 오번역을 방지하였다.

4.5.2 대역어 및 대역 표현 특화

변환 모듈에서 구조 변환 및 어휘 변환을 위해 사용하는 영한 사전 및 어휘패턴의 대역어들에 대한 도메인 특화 작업이 필수적이다. 즉, 기존의 영한 자동번역시스템은 기술논문이나 IT 뉴스와 같은 기술 도메인을 대상으로 하고 있었기 때문에, 이러한 대역어들을 그대로 사용할 경우, 대화체 문장을 번역하는데 있어서 자연스러운 번역이 나오지 않기 때문이다. 이를 위해 GIZA++를 이용한 자동정렬 데이터 및 한국어 코퍼스에서의 대역 어휘 빈도 등을 고려한 대역어 특화 작업을 수행하였다.

4.5.3 대역어 선택

대역어 선택 모호성을 해결하기 위해 변환 모듈에서는 주로 동사와 그 논항의 하위범주정보를 패턴화한 정보를 사용하였다. 이러한 하위범주 정보는 비교적 문장 길이가 짧은 대화체 문장에 있어서 공기어휘보다 높은 성능을 보인다. 현재의 하위범주 정보는 주로 “주어와 동사”, “동사와 목적어”, “형용사와 명사”의 관계를 대상으로 우선적으로 구축되었다.

4.5.4 기타

앞에서 언급하였듯이, 단순히 명사구나 전치사구로만 구성된 문장을 자연스럽게 번역하기 위해서 적당한 종결어미를 할당하여 최종 번역문 생성을 보다 자연스럽게 출력되도록 하였다. 또한 분석 모듈과의 밀결합을 통해, 다양한 숫자 표현을 올바로 변환하고 생성할 수 있는 장치를 마련하였다.

5. 실 험

본 장에서는 제안하는 기술을 기반으로 한 특화된 영한 대화체 자동번역시스템의 성능 평가 결과를 설명한다. 성능 평가를 위해 사용한 테스트셋, 평가 방법, 평가 기준을 설명하면 다음과 같다. 테스트셋은 인터넷에서 수집한 여행 영역의 영어 대화체 200문장을 사용하였으며, 평균 단어수는 5.92 단어이다. 평가 방법은 전문번역가 3인에 의한 원문 대비 정확도를 평균 내어 측정하였다. 표3은 평가 기준을 보인다.

표 3. 번역률 평가를 위한 평가 기준

점수	평가 기준
4.0	원어문의 의미가 그대로 전달된 경우
3.5	복문에서, 문장의 동사구가 정확히 전달되어 문장의 전체적인 의미의 골격이 전달되지만 동사를 제외한 1-2 단어의 대역어가 잘못된 경우
3.0	문장의 동사구가 정확히 전달되어 문장의 전체적인 의미의 골격이 전달되는 경우
2.5	하나의 동사절이라도 정확히 번역되어 부분적으로 문장의 의미를 전달할 경우
2.0	하나 이상의 구가 정확히 번역되지만 전체적인 문장의 의미를 파악하기 어려운 경우
1.0	문장 중에 하나의 단어 또는 구라도 정확히 번역된 경우
0.0	번역문 출력이 안 된 경우

표 4는 번역률 평가 결과를 나타낸다. 번역률은 표 3의 평가 기준에 따라 각 평가 문장에 대해 전문 번역가가 부여한 점수의 평균 점수를 의미한다. 표 4를 통해, 대화체 특화 이전과 비교해서, 새로 구축되고 수정된 번역 지식 등을 알 수 있다.

표 4. 번역률 평가 결과

	특화 전	특화 후
번역메모리 문장수	0	37만 문장
대화체 패턴 구축양	0	2만 패턴
고빈도 어휘 대역어 정련	-	1만7천 어휘
기구축 패턴 대역어 특화	-	3천 패턴
자동번역 엔진 특화	-	모듈별 특화 및 성능 개선
번역률	**72.97%**	**79.7%**

현재의 번역 성능을 보다 개선하기 위해서는 앞에서도 언급하였듯이, 지금까지의 번역 방식인 문장 단위가 아닌 대화 단위의 개선이 필요하며, 이를 위해 엔진의 구성 모듈별로 이를 처리할 수 있는 장치가 마련되어야 한다. 또한 실험에 사용한 문장은 여행자 도메인의 대화체 문장으로 구성되어 있지만, 앞으로 자동번역 수요가 급증할 것으로 보이는 온라인 채팅 등은 또 다른 특성을 지니고 있으며, 많은 어려움을 가지고 있을 것으로 예상된다.

6. 결 론

본 논문에서는 대화체 영한 자동번역을 위해서 대화체 문장이 갖는 특성을 분석하였고, 이를 기반으로 자동번역시스템의 각 모듈 및 지식이 대화체를 위해 특화된 대화체 영한 자동번역시스템에 대해서 설명하였다. 문어체와는 달리 대화체는 아직도 해결하기 어려운 많은 문제점을 가지고 있다. 또한, 대화체의 경우, 문장 단위를 넘어서 대화 단위로의 자동번역도 고려하여야 문맥을 충분히 고려한 올바른 번역 결과를 생성할 수 있다. 따라서, 향후에는 대화체 문맥 처리 기술을 우선적으로 개발하고, 대화 문맥에 기반한 분석 및 변환/생성 기술의 성능 개선이 필요하다.

<참고문헌>

[1] Ying Zhang & Stephan Vogel, PanDoRA, “a large-scale two-way statistical machine translation system for hand-held devices,” *MT Summit XI*, pp.543-550, 2007

[2] Gianni Lazzari, “Speech to speech translation,” *ELRA-HLT Evaluation Workshop*, 110pp, 2005.

[3] 최승권, 김영길, “영한 번역 메모리의 구조화 연구,” *번역학 연구 2009, 10-3호* [게재예정]

[4] 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,” *Proceedings of the 22st Pacific Asia Conference on Language, Information and Computation(PACLIC22)*, pp.161-168, 2008.