구조화된 번역 메모리 기반 영한 메신저 자동 번역 시스템에 관한 연구

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

A Study on English-Korean Messenger MT System based on Structured Translation Memory

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

요 약

본 논문의 목표는 크게 두 가지이다. 하나는 2010년에 개발한 메신저 자동번역 시스템을 소개하는 것이고, 다른 하나는 메신저 대화체 문장을 더욱 고품질로 번역하기 위한 구조화된 번역 메모리(Structured Translation Memory)를 소개하는 것이다. 구조화된 번역 메모리는 기존의 문자열 기반의 번역 메모리와 자동 번역 시스템의 경계를 허무는 개념으로 구조를 표현하는 계층적 번역 메모리들로 구성된다. 구조화된 번역 메모리는 문자열 번역 메모리, 원형 어휘로 구성된 번역 메모리, 고유명사가 청킹된 번역 메모리, 날짜/숫자가 청킹된 번역 메모리, 기본명사구가 청킹된 번역 메모리, 문장 패턴 번역 메모리로 단계적으로 구성된다. 구조화된 번역 메모리를 적용하기 전의 2010년의 영한 메신저 자동 번역 시스템의 번역률이 81.67%였던 반면에, 구조화된 번역 메모리를 적용하려는 2011년의 영한 메신저 자동 번역 시스템의 시뮬레이션 번역률은 85.25%인 것으로 평가되었다. 따라서 구조화된 번역 메모리를 적용하였을 때는 기존의 번역률보다 3.58% 향상할 것으로 예측된다.

1. 서론

영한 자동번역 시스템이 국내에서 개발된 지 어느덧 16년 정도가 되고 있다. 그동안 자동번역 기술의 발전과 더불어 번역 속도와 번역률이 크게 향상되었다. 그 결과 영한 자동번역 시스템의 응용 분야는 특허, 과학 기술 논문[1], IT웹 신문, 기업문서, 군사 매뉴얼 등으로 확장되어 상용화에 성공하였다. 2010년도부터 한국전자통신연구원에서는 영한/한영 메신저 자동번역 시스템을 개발하기 시작했다[2].[1] 앞서 개발한 시스템들이 주로 문어체 위주의 문장을 자동번역 대상으로 삼았다면 메신저 자동번역 시스템은 주로 대화체 문장을 대상으로 한다는 점이 다르다.

본 논문의 목표는 2010년에 개발한 메신저 자동번역 시스템을 소개하고, 메신저 대화체 문장을 더욱 고품질로 번역하기 위한 구조화된 번역 메모리를 소개하는 것이다.

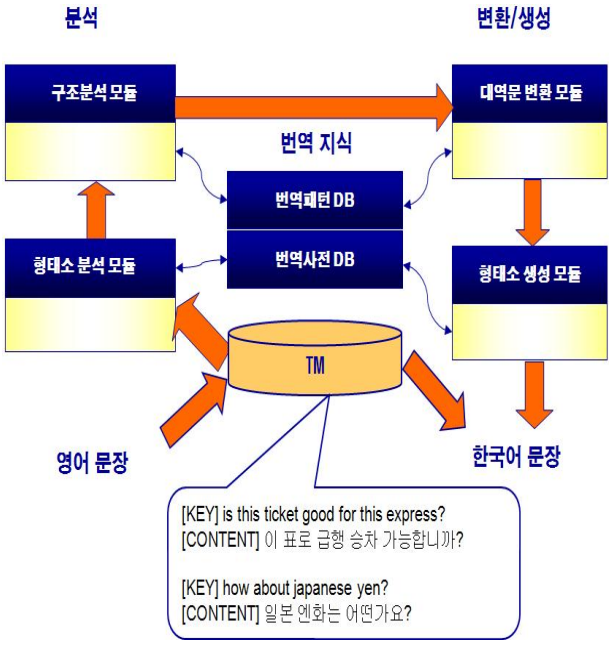
2. 2010년의 메신저 자동번역 시스템

2010년도에 개발한 메신저 자동번역 시스템은 문어체 위주의 패턴기반 영한 자동번역 시스템에 도메인 특화 방법[3]을 적용하여 개발한 시스템이다. 도메인 특화 방법이란 한 도메인으로부터 다른 도메인으로 시스템을 적응해 가는 방법이라고 할 수 있다. 도메인 특화 방법을 적용한 2010년의 영한 메신저 자동번역 시스템의 예가 그림 1이다.

(그림 1) 영한 메신저 자동번역 시스템의 예

1) 본 연구는 지식경제부의 지식경제 기술혁신사업의 일환으로 수행하였음 [2009-S-034, 한중영 대화체 및 기업문서 자동번역 기술 개발].

2010년도의 영한 메신저 자동번역 시스템은 문자열 기반의 번역 메모리(Translation Memory, 이후 TM)를 자동번역 시스템 앞에 부착한 단순한 형태의 구조를 가지는 자동번역 시스템이었다.



(그림 2) 2010년의 영한 메신저 자동번역 시스템 구성도

3. 실험1

2010년의 영한 메신저 자동번역 시스템을 평가하기 위해 영어권 Native speaker와 한국인이 대화한 메신저 대화체 문장 중 100문장을 임의로 추출하였다. 100문장의 평균 단어수는 7.38단어였다. 5인의 번역가에게 평가 점수 부여 기준을 교육한 후, 평가 기준에 따라 각자 평가 점수를 부여하게 하고 각 문장에서 최고/최저 점수를 제외한 3인의 점수에 대한 평균으로 번역률을 계산하였을 때, 수동 번역률은 81.67%가 나왔다.

4. 메신저 대화체 문장의 특징과 2010년 번역기의 문제

문어체 문장과 비교할 때 영어 메신저 대화체 문장의 특징은 대화체 어투, 생략 현상, 슬랭어 표현, 의성어 표현, 반복 기호, 철자 오류, 약어가 자주 나타난다는 것이다. 이러한 대화체 문장의 특징을 해결하기 위해 수행한 도메인 특화 방법은 다음과 같았다.

<표 1> 영어 대화체 문장의 특화

분류	특화 방법	예(특화전->특화후)
어투	번역 패턴 반자동 구축 -의미기반 번역패턴 구축)	Any girls wanna chat?:어떤 여자가 채팅하기를 원합니까?-> 채팅할 여자 누구 없어?
생략	번역엔진모듈 반자동 특화 -고빈도 생략구조 복원	You trying to go home -> You are trying to go home.
슬랭	미등록어 반자동 구축 -슬랭어 사전 구축 -축약 처리	lol->하하하 wb->돌아온 걸 환영해
의성어	번역엔진모듈 반자동 특화 -철자 오류 처리	awwwwwww->aw
대화체 대역어	기구축 용어 반자동 튜닝 -대화체 표현용 사전 대역어 수정	crazy:미친->열광하는
기호	번역엔진모듈 반자동 특화 -기호 및 이모티콘 처리	!!!!!!!!!!!!!!:!!!!!!!!!!!!!!->!
철자오류	번역엔진모듈 반자동 특화 -철자 오류 처리	didnt->didn't
부르기	번역엔진모듈 반자동 특화 -철자 오류 처리	heyyyyy->hey
약어	미등록어 반자동 구축 -약어 사전 구축	NY->New York
고유명사	미등록어 반자동 구축 -고유명사 사전 구축	Buffalo Wings->버팔로윙
대문자	번역엔진모듈 반자동 특화 -소문자 전처리	WHY WOULD ANY ONE...->why would...

이상의 도메인 특화 방법에 의해 메신저 대화체 문장의 특성을 상당한 정도로 해결했음에도 불구하고 몇가지 문제점이 나타났다. 문제점의 주된 원인은 TM의 커버리지였다. 2010년의 번역기에서는 문자열 위주의 TM이 자동 번역이 이루어지기 전에 우선 적용되었다. 이 때문에 입력된 문장이 TM과 매우 유사함에도 불구하고 단순히 문자열 매칭에 의해서만 TM을 활용할 수 있었기 때문에 TM의 커버리지가 낮았었다. 이러한 현상이 다음의 예에서 관찰될 수 있는데 (2)의 영어 원문들이 (1)의 TM을 활용하지 못하여 자동 번역기에 의해 부자연스럽게 자동 번역 되는 것을 알 수 있다:

(1) TM의 예

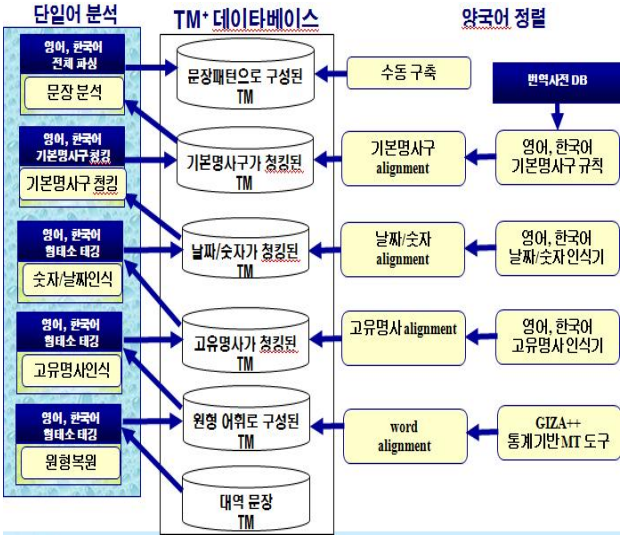
[KEY] Sure no problem. [CONTENT] 네, 물론이죠.
[KEY] One ticket to Paris, please. [CONTENT] 파리 행 표 한 개 주세요.
[KEY] I have $3,000 in cash. [CONTENT] 현금으로 3,000 달러 가지고 있습니다.
[KEY] Could I have two oranges? [CONTENT] 오렌지 2개 주시겠습니까?
[KEY] Have a nice day [CONTENT] 좋은 시간 보내세요

(2) TM과 일치하지 않아 자동 번역된 문장들
Sure, no problem. -> 그럼, 문제없습니다.
One ticket to London, please. -> 런던에게 하나 티켓, 부디.
I have $100 in cash. -> 나는 현금에서 100달러를 가지고 있습니다.
Could I have one ticket? -> 내가 한 티켓을 가질

수 있습니까?
Thank you for choosing San Felice Hotel and have a nice day. -> San Felice Hotel을 선택하여 주셔서 감사하고 좋은 날을 가지고 있으시오.

5. 구조화된 번역 메모리 (TM+)

TM이란 원문의 문장과 그것의 번역된 문장을 하나의 쌍으로 하여 데이터베이스화한 것을 말한다. 이 TM을 사용하는 목적은 번역가가 이전에 번역한 문장이나 반복되는 문장을 중복해 번역하지 않고 번역해 두었던 번역문을 재활용할 수 있도록 하려는 것이다. 이에 반해 구조화된 번역 메모리(Structured Translation Memory, 이후 TM+)는 TM의 재활용률을 높이기 위해 TM에 언어학적인 구조를 반영한 번역 메모리를 말한다. 언어학적인 구조는 원형 어휘, 고유명사 청킹, 날짜/숫자 청킹, 기본명사구 청킹, 문장 패턴을 변수로 치환한 것을 말한다. TM+를 반자동으로 구축하는 방법을 기술하면 다음과 같다.



(그림 3) TM+의 반자동 구축 방법

위의 반자동 구축 방법에 의해 (1)의 예에서 보인 TM을 TM+로 변경한 모습은 다음과 같다.

<표 2> TM+ DB

TM+의 DB	예
원형 어휘로 구성된 TM	[KEY] sure no problem. [CONTENT] 네, 물론이죠.
고유명사가 청킹된 TM	[KEY] one ticket to PRN please [CONTENT] PRN 행 표 한 개 주세요
날짜/숫자가 청킹된 TM	[KEY] i have NUM in cash. [CONTENT] 현금으로 NUM 가지고 있습니다.
기본명사구가 청킹된 TM	[KEY] could i have BNP ? [CONTENT] BNP 주시겠습니까?
문장패턴으로 구성된 TM	[KEY] thank you for VP and have a nice day [CONTENT] VP:어 주어서 감사합니다. 좋은 시간 보내세요

6. TM+ 기반 메신저 자동번역 시스템

TM+기반 영한 메신저 자동 번역 시스템은 기본적으로 패턴기반 자동번역 방법론의 틀에 속한다. 따라서 TM+도 패턴의 연장선 상에 있다고 할 수 있다.

2010년의 TM 연동 메신저 자동번역 시스템과 비교할 때 TM+ 기반 영한 메신저 자동 번역 시스템은 TM 뿐만 아니라 앞서 언급한 계층적으로 쌓여있는 TM+ DB들을 단계적으로 이용한다는 점이 다르다.

TM+기반 영한 메신저 자동 번역 시스템의 개략적인 시스템 구성도는 다음과 같다.

(그림 4) TM+ 기반 메신저 자동번역 시스템 구성도

7. 실험2

실험2는 TM+ 기반 영한 메신저 자동 번역 시스템을 평가한 실험이다. 실험2의 실험 대상은 실험1에서 사용하였던 동일한 평가 문장과 평가 점수를 사용하였다. TM+ 기반 영한 메신저 자동 번역 시스템이 현재 구현 중에 있기 때문에 번역률 측정은 시뮬레이션으로 실시하였다. 즉 이미 구축되어 있는 TM에 TM+가 필요로 하는 변수를 치환한 다음 평가문과 영어 원문에 대해서만 매칭을 실시하여 보았다. 그 결과 다음의 표에서 알 수 있듯이, 2011년도에는 2010년 보다 3.58% 정도가 향상될 것으로 예측되었다.

<표 3> 2011년도 시스템의 번역률 시뮬레이션 결과

	2010년	2011년	비고
TM+가 적용된 문장수		13문장	적용된 TM+의 개수는 20개였음(문장당 1.5개의 TM+가 적용됨)
번역률	81.67%	85.25%	2011년 번역률은 시뮬레이션에 의해 나온 예측 번역률임

<표 3>에 따르면 2011년도 시스템의 번역률 향상은

13문장에 적용된 20개의 TM+에 의해서임을 알 수 있다. 20개의 TM+ 각각이 어떤 효과를 내었는지를 살펴보면 다음의 표와 같았다.

<표 4> 적용된 TM+의 상세 분석

TM+의 종류	적용갯수	적용분포
원형 어휘로 구성된 TM	4	20%
고유명사가 청킹된 TM	2	10%
날짜/숫자가 청킹된 TM	0	0%
기본명사구가 청킹된 TM	5	25%
문장패턴으로 구성된 TM	9	45%
계	20	100%

TM+가 적용된 사례를 살펴보면 다음과 같았다.

(3) 원형 어휘로 구성된 TM에 의해 개선된 예
[원문] I will..... reject it
[2010년 자동번역] 제가 할 수 있을 거예요..... 그것을 거절하시오
[2011년 시물레이션] 그것을 거절하겠어요
[TM+] [KEY] I will reject it [CONTENT] 그것을 거절하겠어요

(4) 고유명사가 청킹된 TM에 의해 개선된 예
[원문] Spanish is my favourite.
[2010년 자동번역] 스페인어는 제 우승 후보입니다.
[2011년 시물레이션] 스페인어는 내가 좋아하는 것입니다
[TM+] [KEY] PRN is my favourite [CONTENT] PRN:는 내가 좋아하는 것입니다

(5) 기본명사구가 청킹된 TM에 의해 개선된 예
[원문] Seoul street never turns lights
[2010년 자동번역] 서울 길은 결코 불을 바꾸지 않습니다
[2011년 시물레이션] 서울 길은 결코 불을 키지 않습니다
[TM+] [KEY] BNP never turn light [CONTENT] BNP:는 결코 불을 키지 않습니다

(6) 문장패턴으로 구성된 TM에 의해 개선된 예
[원문] It was sooooooo cute, you know?
[2010년 자동번역] 그것은 정말 귀여웠는지 당신이 압니까?
[2011년 시물레이션] 그것은 정말 귀여웠습니다 알죠?
[TM+] [KEY] S! , you know? [CONTENT] S! 알죠?

8. 결론

본 논문의 목표는 크게 두 가지였다. 하나는 2010년에 개발한 메신저 자동번역 시스템을 소개하는 것이었고, 다른 하나는 메신저 대화체 문장을 더욱 고품질로 번역하기 위한 구조화된 번역 메모리(Structured Translation Memory, 본문에서는 TM+로 명명하였음)를 소개하는 것이었다. TM+란 번역 메모리(Translation Memory, 본문에서는 TM으로 명명하였음)를 포함하는 계층적 TM을 말하는 것으로 1) 원형 어휘로 구성된 TM 2) 고유명사로 청킹된 TM 3) 날짜/숫자가 청킹된 TM 4) 기본명사구가 청킹된 TM 5) 문장패턴으로 구성된 TM으로 구성된다. 따라서 TM+는 기존의 TM과 자동 번역의 경계를 허무는 개념이라고 할 수 있다.

영어권 Native speaker로부터 직접 수집한 메신저 대화체 문장을 대상으로 번역률을 평가한 결과, 2010년의 영한 메신저 자동 번역 시스템은 81.67%였던 반면 TM+를 적용하려는 2011년의 영한 메신저 자동 번역 시스템은 85.25%였다. 따라서 TM+를 적용할 경우 기존의 번역률보다 3.58% 향상할 것으로 예측된다.

참고문헌

[1] 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.
[2] 최승권, 이기영, 노윤형, 권오욱, 김영길. "문어체에서 대화체 문장 패턴기반 영한 번역기로의 특화", 한글 및 한국어 정보처리, 2010, pp.136-140.
[3] Zajac R. "MT Customziation". In Machine Translation Summit Workshop. 2003.