번역학연구
2009년 가을 제10권3호

영한 번역 메모리의 구조화 연구

최승권 · 김영길
(한국전자통신)

1. 서론

번역 메모리(Translation Memory, 이후 TM)란 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. 따라서 TM을 사용하게 되면 번역가는 번역의 일관성과 동시에 번역의 효율성을 높일 수 있는 장점을 가질 수 있다. 이와 같은 TM의 장점에도 불구하고 번역가 지원 도구(Computer-Aided Translation tool, 이후 CAT)[1]에서 TM이 활발히 활용되지 못하는 이유는 TM의 재활용률이 낮기 때문이다. 왜냐하면 TM은 단순히

1) CAT의 대표적인 제품으로는 Trados, Wordfast, SDL Trados2006, Deja Vu 등이 있으며, CAT는 TM 이외에도 데이터 관리 도구, 용어 관리 도구, 문서 처리 도구 등을 포함한다(박주형 외23)

문자열(string)로만 기술되기 때문에 한 글자만 틀려도 해당 TM을 찾을 수 없기 때문이다. 따라서 번역가들이 CAT를 더욱 활발히 사용하기 위해서는 TM의 재활용률을 높이는 것이 무엇보다 중요하다(Lagoudaki 25). 이와 관련하여 본 논문은 두 가지 목표를 가진다. 하나는 문자열 위주의 TM이 갖는 낮은 재활용률을 높이기 위해 언어학적 구조가 부여된 구조화된 번역 메모리(Structured Translation Memory, 이후 TM^+)를 개발하는 것과, 다른 하나는 TM^+와 영한 자동 번역 시스템을 연동하여 번역 품질이 개선되는 것을 확인하는 것이다.[2)]

본 논문의 구성은 다음과 같다. 2장에서는 TM의 재활용 과정에서 나타나는 문제점을 영어-한국어 예문을 통해 살펴본 후 기존의 연구 사례를 통하여 문제점의 해결 방식을 검토하고자 한다. 3장에서는 2장에서의 관찰을 토대로 TM^+의 구조를 기술할 것이다. 4장에서는 TM^+를 영한 자동 번역 시스템과 연동했을 때 자동 번역 시스템의 번역률에 끼치는 영향을 실험을 통해 살펴보고자 한다. 5장에서는 전체적인 결론을 내리고자 한다.

2. TM 재활용의 문제점과 그 해결을 위한 기존의 연구들

2.1. TM 재활용의 문제점

TM을 활용하는 데 있어서 낮은 재활용률의 가장 큰 원인은 TM이 문자열로만 구성되어 있다는 데 있다. 따라서 문자열로 되어 있는 TM을 더 재활용할 수 있는 방법을 찾는 것이 무엇보다 중요하다고 할 수 있다. TM 재활용의 문제점 중 첫 번째 원인으로 꼽을 수 있는 것은 어휘의 축약형과 기본형이 공존하는 경우라고 할 수 있다. 예문 (1)이 그것을 보여준다.

(1) [입력문] I'm not feeling well.
[TM] I am not feeling well. <-> 컨디션이 별로 안 좋아요.

---

2) 본 논문에 대해 정확하면서도 세밀하게 심사해 주신 세분의 익명의 심사위원께 감사의 말씀을 드리며, 본 논문에서 나오는 데이터 오류나 논문 전개상의 논리의 부정확함이 있다면 순전히 저자의 책임임을 밝힙니다.

(1)의 예는 번역할 문장인 [입력문]이 “I'm”과 같이 축약형의 어휘를 가지고 있어서 번역 메모리인 [TM]에 기본형인 “I am”의 문장이 존재하더라도 일치하지 않아 TM을 재활용하지 못하는 경우이다. 두 번째 원인으로는 구두점의 있고 없음에 따른 불일치의 문제이다.

(2) [입력문] Sure, no problem.
[TM] Sure no problem. <−> 네, 물론이죠.

(2)의 예는 [입력문]이 “Sure,”과 같이 콤마를 포함하고 있어 콤마를 포함하지 않은 [TM]의 영어 문장과 일치하지 않아 TM을 재활용하지 못하는 경우이다. 세 번째 경우는 문장 서두에 나타나는 감탄사, 긍정/부정어, “I‘m sorry, but”과 같은 문두 부사 상당 어구로 인한 불일치의 경우다.

(3) [입력문] I'm sorry, but we are out of tomato juice.
[TM] We are out of tomato juice. <−> 토마토 주스가 다 떨어졌어요.

(3)의 예는 [입력문]이 “I’m sorry, but”과 같이 문두 부사 상당 어구를 포함하고 있어 문두 부사 상당 어구를 포함하지 않은 [TM]의 영어 문장과 일치하지 않아 TM을 재활용하지 못하는 경우이다. 네 번째 경우는 인명, 지명, 회사명과 같은 고유 명사가 일치하지 않아 나타나는 현상이다.

(4) [입력문] One ticket to London, please.
[TM] One ticket to Paris, please. <-> 파리 행 표 한 개 주세요.

예문 (4)는 [입력문]의 “London”과 같은 고유명사가 [TM]의 “Paris”와 같은 고유명사와 일치하지 않아 TM이 적용되지 않는 예이다. 다섯 번째는 날짜, 전화번호, 금액 등과 같은 숫자가 일치하지 않아 TM이 적용되지 않는 경우이다.

(5) [입력문] I have $100 in cash.
[TM] I have $3,000 in cash. <−> 현금으로 3,000 달러 가지고 있습

니다.

예문 (5)는 [입력문]의 금액을 나타내는 숫자 “$100”가 [TM]의 “$3,000”과 일치하지 않아 TM이 적용되지 않은 경우다. 여섯 번째는 기본 명사구가 일치하지 않아 TM이 적용되지 않는 경우이다.

(6) [입력문] Could I have one ticket?
[TM] Could I have two oranges? <−> 오렌지 2개 주시겠습니까?

예문 (6)은 [입력문]과 [TM] 사이에 “one ticket”과 “two oranges”의 명사구에서만 차이가 발생해 TM이 적용되지 않은 경우다. 일곱 번째는 숙어 표현에 관한 경우다.

(7) [입력문] Thank you for choosing San Felice Hotel and have a nice day.
[TM] Have a nice day <−> 좋은 시간 보내세요
Thank you for ...<−> ... 어 주셔서 감사합니다

예문 (7)에서 [입력문]의 “thank you for”와 “have a nice day”가 [TM]에 각각 존재하지만 “choosing San Felice Hotel” 때문에 [TM]이 적용되지 않는 경우다.

이상의 문자열로 된 기존 TM의 문제점을 요약하면 다음과 같다. 1) 입력문의 어휘가 축약형 형태로 되어 있어 기본형의 어휘를 포함하는 TM을 활용하지 못함 2) 입력문이 구두점을 포함하고 있어 구두점이 없는 TM을 활용하지 못함 3) 입력문이 문두 부사 상당 어구를 포함하고 있어 문두 부사 상당 어구를 포함하지 않는 TM을 활용하지 못함 4) 입력문이 고유 명사를 포함하고 있어 다른 고유명사를 포함하고 있는 TM을 활용하지 못함 5) 입력문이 숫자를 포함하고 있어 다른 숫자를 포함하고 있는 TM을 활용하지 못함 6) 입력문의 기본 명사구 때문에 다른 기본 명사구를 가지고 있는 TM을 활용하지 못함 7) 입력문의 숙어를 제외한 구 때문에 TM을 활용하지 못함. 따라서 이와 같은 기존의

TM 문제점을 해결하는 것이 본 논문의 목표이다.

2.2. 기존의 연구들

기존의 연구에서는 TM의 재활용률을 높이기 위해 TM에 언어학적인 구조를 반영하려고 하였다. 언어학적인 구조를 어디까지 확장하여 반영하느냐에 따라 그동안의 연구를 정리하면 다음과 같다.

2.2.1. 어휘소까지 확장하는 방법

TM의 재활용률을 높이기 위해 행한 초기의 방법이 어휘소(lexeme)까지 확장하는 방법이다(Carl et.al. 617). 어휘소까지 확장하는 방법은 말 그대로 문자열 기반의 TM을 어휘소 기반의 TM으로 추상화하여 재활용률을 높이는 방법이다. 이 연구에 따르면 어휘소 기반의 TM은 문자열 기반의 TM보다 재활용률은 높지만 번역된 결과의 품질은 떨어진다는 결론에 도달하였다.

2.2.2. 품사까지 확장하는 방법

어휘소보다 재활용률을 더욱 높이기 위해 고안한 방법이 품사까지 확장하는 방법이다(Rapp 466)(Planas et.al. 332). 이 방법은 형태소 분석기를 이용해 원문의 어휘에 품사를 부여하고 문자열, 어휘소, 품사를 점차적으로 모두 활용하는 방법이다. Rapp(2002, 466)에 따르면 독일어 문장 “Dann bereitete er das Essen”에 대한 영어 번역문을 찾기 위해 형태소 분석을 하게 되는데 형태소 분석 결과는 “Dann/adverb bereitet/verb er/pronoun das/article Essen/noun”과 같다. 독일어 문장의 품사열에 대응되는 영어 품사열은 영어에서는 동사가 대명사 다음에 나오기 때문에 “adverb pronoun verb article noun”가 된다. 이 품사열에 대응되는 영어 후보들을 대상으로 독영 사전 (8)의 예에 의해 영어 후보들을 나열하면 예 (9)와 같다.

(8) 독영 사전

German	English
dann	then (adverb)
bereitete	prepared (verb), caused (verb)
er	he (pronoun)
das	the (article)
Essen	meal (noun), food (noun), Essen (proper noun)

(9) [KEY] adverb_pronoun_verb_article_noun
[CONTENT] then he prepared the meal
then he caused the meal
then he prepared the food
then he caused the food
then he prepared the Essen.
then he caused the Essen.

위의 영어 후보들을 대상으로 구조 분석 모호성, 공기 정보(co-occurrence information)에 의한 의미 모호성을 해결하면 “Then he prepared the meal”을 만들 수 있다는 것이다.

2.2.3. 품사 이상의 구조까지 확장하는 방법

TM의 재활용률을 높일 뿐 아니라 예제기반 자동 번역 시스템 (Example-based machine translation system) (Hutchins 63)과 연동하기 위해 TM은 어휘소와 품사를 넘어 더 추상적인 언어학 구조까지 발전하였다. 언어학 구조는 구 단위에서 절 단위로 확장되었으며, 구축 방법은 병렬 말뭉치를 대상으로 수동 구축에서 통계적 자동 구축으로 발전하였다. 구 단위 중에 명사구를 TM에 반영하는 것이 우선 개발되었다(Vogel et.al. 1132)(Hodasz et.al. 82). 다음의 예가 명사구까지 확장한 독일어-영어 번역 메모리의 예이다.

(10) # SURNAME am Apparat # this is SURNAME speaking # -3.3

```
# NP dauert DATE          # NP takes DATE              # -3.3
# nehmen PPER NP DATE # let PPER take NP DATE     # -4.6
```

(10)의 예에서 “NP”가 명사구를 나타내는 변수이며, 품사를 위한 변수로서 “SURNAME”, “DATE”, “PPER”은 각각 이름의 성, 날짜, 인칭 대명사를 나타내는 변수이다. 맨 뒤의 “-4.6”은 점수를 나타낸다. 명사구를 넘어서 더 큰 언어학 구조로 확장한 경우는 구 사전(phrasal lexicon)(Schäler 49)과 하위절(Simard et.al. 335)까지 확장한 경우가 있었다. 구 사전 방법은 병렬 말뭉치로부터 대응되는 표현들을 통계적으로 자동 수집하여 두었다가 입력문이 들어오면 입력문과 일치하는 구 사전 표현들을 찾아서 치환하는 방법이다.

【그림 1】 구 사전으로 기술된 TM

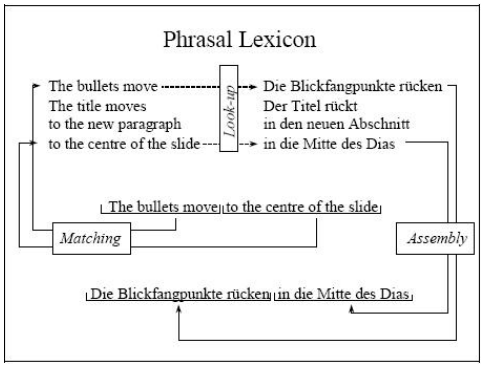


그림 1은 영어-독일어 병렬 말뭉치를 대상으로 구축된 구 사전의 예로써 “The bullets move to the center of the slide”이라는 문장이 입력되었을 때, 구 사전에 저장되어 있던 “The bullets move → Die Blickfangpunkte rűcken”과 “to the center of the slide → in die Mitte des Dias”을 가져와서 부분 번역을 한다는 것을 말한다.

【그림 2】 하위절까지 확장한 TM

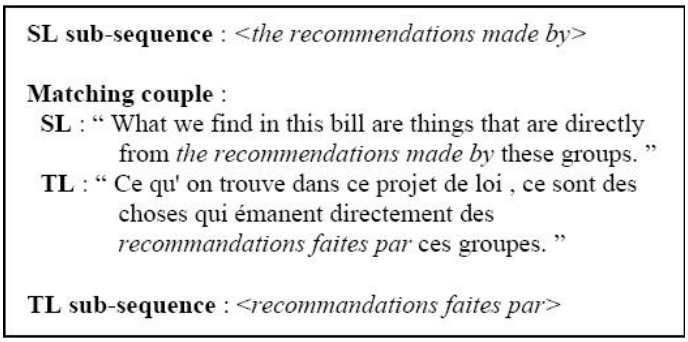


그림 2도 구 사전의 구축과 유사하게 병렬 말뭉치로부터 연속적으로 일치하는 절 단위 표현들을 통계적으로 수집하여 번역시에 활용한다는 의미이다.

이상의 기존 연구를 요약하면 다음과 같다. TM의 재활용률을 높이기 위해 기존의 연구는 어휘소, 품사, 명사구, 구 사전, 하위 절까지 확장하는 방법을 시도한 바 있다. 그러나 이 모든 방법들은 영어-독일어나 영어-스페인어와 같이 구조가 유사한 어족들에만 적용되었지 영어-한국어와 같은 구조가 다른 어족에는 적용되지 못했다. 그 이유는 영어-한국어와 같은 번역문에 대한 단어 정렬(alignment) 방법이 아직 구조가 다른 어족에는 정확하지 않기 때문이다.

3. 구조화된 번역 메모리(TM$^+$)의 반자동 구축 방법

TM$^+$는 TM의 재활용 문제점을 해결하면서도 기존 연구의 장점을 수용하는 것을 첫 번째 목표로 하며 또한 패턴 기반 자동번역 시스템과 자연스럽게 연동함으로써 궁극적으로는 직역에 머물러 있는 현재의 영한 자동 번역 시스템의 번역 품질을 의역 수준까지도 끌어올리는 것을 두 번째 목표로 하고 있다. 2장에서 언급한 TM 재활용의 문제점과 이를 해결하기 위한 기존의 연구들을 포함하는 TM$^+$를 반자동으로 구축하는 방법은 다음과 같다.

【표 1】 TM+의 반자동 구축 방법

TM 재활용의 문제점	TM+의 반자동 구축 방법
어휘의 축약형으로 인한 불일치	전처리
구두점으로 인한 불일치	
문두 부사 상당 어구로 인한 불일치	문두 부사 상당 어구 제거 및 확장
고유명사로 인한 불일치	고유 명사 청킹 및 PRN 치환
날짜, 전화번호, 금액과 같은 수로 인한 불일치	숫자 청킹 및 NUM 치환
기본명사구로 인한 불일치	기본 명사구 청킹 및 BNP 치환
문장내 숙어적 표현으로 인한 불일치	숙어 청킹 및 나머지 부분 변수 치환

표 1에서 청킹(chunking)은 해당 정보를 묶는다는 의미로써 고유 명사 청킹은 고유 명사와 관련된 정보를 묶는 것을 의미한다. TM+를 이루는 각 단계와 그것을 반자동으로 구축하는 방법을 2장에서 이미 제시한 TM의 예문을 가지고 아래의 각 절에서 상세히 기술하도록 하겠다.

3.1. 전처리

TM 재활용률을 높이기 위해 우선 해야 할 일은 TM의 변이 형태를 줄이는 것이다. 즉 TM을 구성하는 문장 성분들의 변이 형태를 대표 형태로 통일하는 것이다. 언어에 따라 전처리 대상이 되는 것은 다양하겠으나 영어를 대상으로 하였을 때 축약 어휘의 복원과 구두점을 제거하는 것이 전처리의 대상이 된다.

(11) 축약 어휘 복원의 예
[TM] I'm not feeling well. <-> 컨디션이 별로 안 좋아요.
[TM+] I am not feeling well. <-> 컨디션이 별로 안 좋아요.

(12) 구두점 삭제의 예
[TM] Sure, no problem. <-> 네, 물론이죠.
[TM+] Sure no problem. <-> 네, 물론이죠.

3.2. 문두 부사 상당 어구 제거 및 확장

문장 전체의 의미에는 영향을 끼치지 않으면서 문두 부사 상당 어구가 앞 문장과 자연스럽게 연결하기 위해 사용되는 경우가 있다. 이러한 문두 부사 상당 어구를 TM에서 제거한 나머지 문장도 TM으로 재활용하는 방법이 문두 부사 상당 어구 제거 및 확장 방법이다.

(13) 문두 부사 상당 어구 제거 및 확장된 예
[TM] I'm sorry, but we are out of tomato juice. <–> 미안하지만, 토마토 주스가 다 떨어졌어요.
[TM^{+}] I am sorry but we are out of tomato juice <–> 미안하지만, 토마토 주스가 다 떨어졌어요.
We are out of tomato juice <–> 토마토 주스가 다 떨어졌어요.

예 (13)에서 “I'm sorry, but we are out of tomato juice <–> 미안하지만, 토마토 주스가 다 떨어졌어요.”라는 TM을 대상으로 전처리된 “I am sorry but” 과 “미안하지만,”을 삭제한 “we are out of tomato juice <–> 토마토 주스가 다 떨어졌어요.”를 TM에 추가하여 TM을 확장하는 방법이 문두 부사 상당 어구 제거 및 확장 방법이다. 문두 부사 상당 어구를 제거하는 것은 수동으로 구축된 문두 부사 상당 어구 목록에 의해 반자동으로 이루어질 수 있다.

3.3. 고유 명사 청킹 및 PRN 치환

인명, 지명, 회사명과 같은 고유 명사만 일치하지 않아서 TM을 활용하지 못하는 경우가 있다. 이런 경우를 방지하기 위해 고유 명사를 변수로 치환하여 TM의 재활용성을 높이는 것이 고유 명사 청킹 및 PRN 치환이다. 다음의 예에서 “Paris”를 변수 “PRN”으로 치환하여 TM과 함께 TM^{+}에 등록하는 것이다. 물론 한 단어 고유 명사뿐만 아니라 “New York”과 같은 두 단어 이상도 해당한다.

(14) 고유 명사 청킹 및 PRN 치환의 예
[TM] One ticket to Paris, please. <–> 파리 행 표 한 개 주세요.

[TM+] One ticket to Paris please. <-> 파리 행 표 한 개 주세요.
One ticket to PRN please <-> PRN 행 표 한 개 주세요

고유 명사를 PRN으로 치환하는 방법은 수동으로 구축된 고유 명사 목록에 의해 반자동으로 이루어질 수 있다.

3.4. 숫자 청킹 및 NUM 치환

날짜, 전화번호, 경제 수치 등과 같은 수 표현의 차이로 인해 TM이 적용되지 않는 경우 이러한 수를 변수 “NUM”으로 치환하는 것이 숫자 청킹 및 NUM 치환 방법이다. 다음의 예에서 “$3,000”를 변수 “NUM”으로 치환하고 “$3,000”에 대응되는 한국어 표현 “3,000달러”를 “NUM”으로 치환하여 만든 것이 TM+이다.

(15) 숫자 청킹 및 NUM 치환
[TM] I have $3,000 in cash. <-> 현금으로 3,000달러 가지고 있습니다.
[TM+] I have $3,000 in cash. <-> 현금으로 3,000달러 가지고 있습니다.
I have NUM in cash. <-> 현금으로 NUM 가지고 있습니다.

숫자 표현을 NUM으로 치환하는 방법도 숫자 목록를 수동으로 구축하고 이것을 토대로 반자동으로 이루어진다.

3.5. 기본 명사구 청킹 및 BNP 치환

기본 명사구란 고빈도로 자주 나타나며 명사구 중 가장 기초적인 명사구라고 할 수 있다. 영어에서 “명사”, “관사 명사”, “형용사 명사”, “수사 명사”, “관사 형용사 명사” 등과 같은 경우이다. 이런 기본 명사구만의 차이로 TM을 활용하지 못하는 경우 기본 명사구를 변수 “BNP”로 치환하여 TM의 활용성을 높이는 것이 기본 명사구 청킹 및 BNP 치환의 목적이다. 다음의 예에서 “two oranges”를 변수 “BNP”로 치환하고 “two oranges”에 대응되는 한국어의 “오렌지 2개”를 “BNP”로 치환하여 만든 것이 TM+이다.

(16) 기본명사구가 청킹된 TM (BNP-TM)
[TM] Could I have two oranges? <-> 오렌지 2개 주시겠습니까?
[TM+] Could I have two oranges? <-> 오렌지 2개 주시겠습니까?
Could I have BNP ? <-> BNP 주시겠습니까?

기본 명사구를 BNP로 치환하는 방법은 자동 번역용 전자사전과 기본 명사구 치환 규칙을 수동으로 구축하고 이것을 토대로 반자동으로 이루어지는 것이다.

3.6. 숙어 청킹 및 나머지 부분 변수 치환

TM+에서 가장 난이도가 높은 TM이 숙어 청킹 및 나머지 부분 변수 치환이다. 이 방법은 숙어 부분을 제외한 나머지 부분에 적절한 구나 절 변수를 표시하여 숙어 부분을 재활용하는 방법이다. 변수는 동사구이면 “VP”를 문장이면 “S”를 부사절이면 “SBAR”를 부여하도록 하였다. (17)의 예에서 “thank you for”와 “have a nice day”가 반복적으로 사용되었기 때문에 숙어 표현으로 간주된 것이고 “choosing San Felice Hotel”은 교체되어도 가능한 부분이기 때문에 변수로 치환한 것이다.

(17) 숙어 청킹 및 나머지 부분 변수 치환 예
[TM] Thank you for choosing San Felice Hotel and have a nice day. <-> San Felice 호텔을 선택해 주셔서 감사합니다. 좋은 시간 보내세요
[TM+] Thank you for choosing San Felice Hotel and have a nice day <-> San Felice 호텔을 선택해 주셔서 감사합니다. 좋은 시간 보내세요
Thank you for VP and have a nice day <-> VP:어 주어서 감사합니다. 좋은 시간 보내세요

(17)의 예에서 “choosing San Felice Hotel”은 동사구에 해당하므로 “VP” 변수로 치환한 것이고 “VP”에 대응되는 한국어 표현 “San Felice 호텔을 선택해”는 “VP”로 치환하고서 변수 “VP”와 한국어 어미인 “어”를 구분하기 위해

":"를 표시하였다.

이상의 6단계에 이르는 TM^+ 구축 방법에 의해 임의의 두 문장을 TM^+로 만들어 보면 다음과 같다.

(18) Two economy class seats to Los Angeles, please. <-> 로스앤젤레스행 보통석으로 두 자리 주세요.
[TM^+] Two economy class seats to Los Angeles please <-> 로스앤젤레스행 보통석으로 두 자리 주세요. (전처리)
NUM economy class seats to PRN please <-> PRN 행 보통석으로 NUM 자리 주세요. (PRN 치환 -> NUM 치환)

(19) Yes, a surfboard costs twenty-five dollars for two hours and forty dollars for a full day <-> 예, 서프보드는 두 시간에 25 달러이고 하루 종일은 40 달러예요
[TM^+] yes a surfboard costs twenty-five dollars for two hours and forty dollars for a full day <-> 예, 서프보드는 두 시간에 25 달러이고 하루 종일은 40 달러예요 (전처리)
a surfboard costs twenty-five dollars for two hours and forty dollars for a full day <-> 서프보드는 두 시간에 25 달러이고 하루 종일은 40 달러예요 (문두 부사 상당 어구 제거 및 확장)
yes BNP costs NUM dollars for NUM2 hours and NUM3 dollars for BNP2 <-> 예, BNP:는 NUM2 시간에 NUM 달러이고 BNP2:은 NUM3 달러예요. (NUM 치환 -> BNP 치환)
BNP costs NUM dollars for NUM2 hours and NUM3 dollars for BNP2 <-> BNP:는 NUM2 시간에 NUM 달러이고 BNP2:은 NUM3 달러예요. (NUM 치환 -> BNP 치환)

예 (18)은 "Two economy class seats to Los Angeles, please. <—> 로스앤젤레스행 보통석으로 두 자리 주세요."라는 영한 대역 문장을 TM^+로 만들면 "전처리", "고유 명사 청킹 및 PRN 치환", "숫자 청킹 및 NUM 치환"의 단계

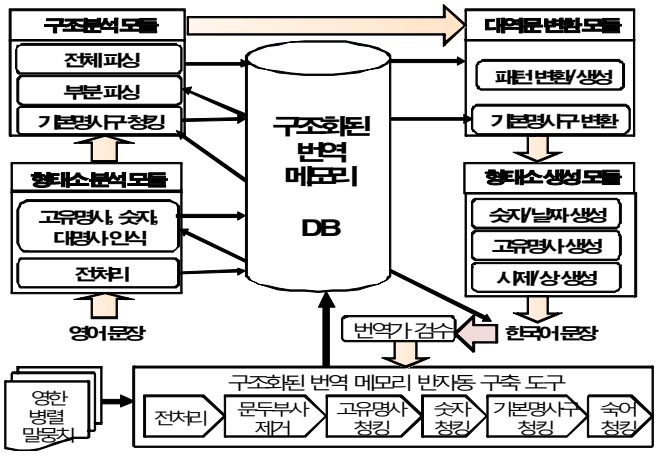
를 거쳐 “Two economy class seats to Los Angeles please <-> 로스앤젤레스 행 보통석으로 두 자리 주세요.”과 “NUM economy class seat to PRN please <-> PRN 행 보통석으로 NUM 자리 주세요.”의 두 개의 TM이 TM+에 저장된다는 것을 말한다. 예 (19)은 “전처리”, “문두 부사 상당 어구 제거 및 확장”, “숫자 청킹 및 NUM 치환”, “기본 명사구 청킹 및 BNP 치환”에 의해 “Yes a surfboard costs twenty-five dollars for two hours and forty dollars for a full day <-> 예, 서프보드는 두 시간에 25 달러이고 하루 종일은 40 달러예요.”는 위와 같이 4 종류의 TM+로 만들어진다는 것을 말한다. BNP2 또는 NUM2, NUM3 등은 동일한 변수가 나타날 때 변수를 구분하기 위한 표현이다.

4. TM+의 효과 실험

4.1. TM+ 기반 시스템 구성도

TM+와 연동된 영한 자동 번역 시스템의 개략적인 시스템 구성도는 다음과 같다.

【그림 3】 TM+가 반영된 영한 자동 번역 시스템 구성도



시스템 구성도는 크게 두 부분으로 나눌 수 있다. 하나는 TM+를 반자동으로 구축하는 “구조화된 번역 메모리 반자동 구축 도구”이며 다른 하나는 입력문에 대해 TM+를 기반으로 자동 번역을 하는 “구조화된 번역 메모리 기반 자동 번역 시스템”이다. “구조화된 번역 메모리 반자동 구축 도구”에서는 대량의 영한 병렬 말뭉치를 대상으로 전처리, 문두 부사 상당 어구 제거, 고유 명사 청킹, 숫자 청킹, 기본 명사구 청킹, 숙어 청킹에 의해 “구조화된 번역 메모리 DB”를 대량으로 반자동 구축할 수 있다. “구조화된 번역 메모리 기반 자동 번역 시스템”에서는 입력된 영어 문장이 전처리된 번역 메모리와 일치하면 전처리된 번역 메모리의 한국어 문장이 출력하게 되고, 일치하지 않는다면 상위 단계로 이동한다. 상위 단계에서는 고유명사, 숫자, 대명사, 기본 명사구가 변수로 치환된 번역 메모리와 비교하게 되며, 일치하면 변수에 대한 변환 및 생성에 의해 한국어 문장이 출력되고, 일치하지 않으면 문장이 구조 분석된다. 문장의 구조 분석을 담당하는 파싱(Parsing)에 의해 숙어가 인식되고 숙어를 제외한 구는 구 단위 번역 메모리에 의해 자동 번역되는 것이다. 마지막으로 번역가에 의해 번역되지 않고 자동 번역된 결과는 “번역가 검수”에 의해 검토가 이루어진 후 다시 구조화된 번역 메모리 절차를 거쳐 DB에 저장되게 된다.

4.2. 번역률 실험

TM+의 효과를 검증하기 위해 번역률 실험을 하여 보았다. 번역률 실험은 TM과 자동 번역이 연동된 결과와 TM+와 자동 번역이 연동된 결과로 나누어 수행되었다. 번역률 실험을 위해 사용한 영한 대역 문장은 여행자용 자동통역 시스템을 위해 구축되어 있던 93,161 문장이었다. 평가문은 인터넷을 통해 관광과 관련한 영어 대화체 문장 200문장을 임의로 수집하여 만들었다. TM+에 사용된 93,161 문장은 중복되는 문장이 하나도 없도록 만들어진 것으로 1 문장당 평균 단어수가 6.9 단어였으며 평가문 200문장은 1 문장 당 평균 단어수가 5.9 단어였다. 다음은 각각 93,161 문장 중 일부와 평가문의 일부를 보여 준다.

(20) TM+에 사용된 93,161 문장 중 일부

000290 Hamburger and salad please <-> 햄버거와 샐러드 주세요.

000291 Which salad <-> 어떤 샐러드로 하실래요.
000292 On this side <-> 이쪽에 있는 걸로 주세요.

(21) 평가문에 사용된 일부

What time does the next train to London leave?
At 16:35, from platform 8.
Is it a direct train to London?
No, you have to change trains at Birmingham.
I see.
One ticket to London, please.
Single or return, sir?

평가문을 대상으로 번역률 평가를 수행하기 위해 수동 평가 기준과 평가 방법을 설정하였다. 수동 평가 기준은 원문의 의미가 번역문에 얼마나 충실히 전달되었는가에 초점을 두고 만들었으며 LREC(Language Resources and Evaluation Conference)에서 사용한 기준표를 토대로 만들었다. 또한 수동 평가 방법은 전문 번역가들이 평가한 결과에 대한 평균값에 의해 만들어졌다.

【표 2】 수동 평가 점수 부여 기준

점수	기 준
4.0	원어문의 의미가 그대로 전달된 경우
3.5	원문의 문장 전체가 잘 분석되어 문장의 전체적인 의미의 골격이 전달되지만 동사를 제외한 1-2단어의 대역어가 잘못된 경우
3.0	원문의 문장 전체가 잘 분석되어 문장의 전체적인 의미의 골격이 전달되지만 여러 단어의 대역어가 잘못된 경우
2.5	원문의 문장 전체의 분석은 실패했으나, 하나 이상의 동사구가 잘 분석되고 정확히 번역되어 부분적으로 문장의 의미가 전달된 경우
2.0	원문의 문장 전체의 분석은 실패하여 전체적인 문장의 의미를 파악하기 어려우나, 하나 이상의 명사구가 잘 분석되고 정확히 번역됨.
1.0	원문의 문장 전체의 분석은 실패하여 전체적인 문장의 의미를 파악하기 어려우나, 문장 중에 하나 이상의 단어 또는 한 개의 명사구라도 정확히 번역된 경우
0.0	원문이 번역문에 그대로 출력됨

(22) 수동 평가 방법

3인의 번역가에게 '수동 평가 점수 부여 기준'을 교육한 후, 평가를 실시하고 3인의 평균값으로 평가를 함.

번역률(%) = $(\sum_{i=1}^{n}(\sum_{j=1}^{3}(score_j/4))/3)/n \times 100.0$,

(여기서 n은 추출된 평가문의 수이며, $score_j$는 j번째 전문번역가에 의해 평가된 점수를 말한다)

4.2.1. 기존의 TM 평가 (Baseline 평가)

문자열로만 구성된 기존의 TM 측정은 표층 형태 그대로 평가되었다. 그 결과는 다음과 같았다.

【표 3】 문자열로만 구성된 TM의 적용 전과 적용 후의 번역 품질 비교

	TM 적용 전	TM 적용 후
TM수		93,161 문장
평가문장수	200문장	200문장
적용문장수		15 문장
번역률	72.97%	74.44%

위의 도표로부터 알 수 있는 것은 TM이 적용되기 전의 번역률은 72.97%였는데 TM을 적용함으로써 15 문장이 일치하여 74.44%가 되었다는 것이다. 다음의 예는 TM을 적용하기 전과 후의 구체적인 실례이다.

(23) a. Here you are.

[적용 전] 여기에서 당신은 있다. (점수: 2.3)

[적용 후] 여기 있습니다. (점수: 4.0)

b. I wouldn't count on it.

[적용 전] 나는 그것에 의지하지 않은 것이다. (점수: 3.3)

[적용 후] 전 그거 기대 안 해요. (점수: 4.0)

c. I'll be right with you.

[적용 전] 나는 당신의 마음에 들 것이다. (점수: 3.0)

[적용 후] 곧 가겠습니다. (점수: 4.0)

d. Please wait to be seated.

[적용 전] 앉기를 부디 기다리시오. (점수: 2.8)

[적용 후] 잠시만 기다리시면 안내해 드릴께요. (점수: 4.0)

4.2.2. TM^{+} 평가

TM^{+} 평가는 세 가지로 나누어 평가하였다. 첫 번째는 전처리 및 문두 부사 상당 어구와 관련된 평가였으며, 두 번째는 고유명사, 숫자, 기본명사구가 변수로 치환된 경우의 평가였으며 마지막으로 숙어 관련 평가였다. TM과 TM^{+}을 비교하면 다음과 같았다.

【표 4】 TM과 TM^{+}의 적용 후 번역 품질 비교

	TM 적용 후	TM^{+} 적용 후			
		전처리 TM 적용 후	PRN/NUM/BNP TM 적용 후	숙어 TM 적용 후	누적수
TM수	93,161 문장	91,392 문장	58,744 문장	-	150,136 문장
평가문장수	200 문장	200 문장			
적용문장수	15 문장	37 문장	36 문장	3 문장	76 문장
번역률	74.44%	79.78%	82.03%	82.40%	82.40%

표 4로부터 우선 알 수 있는 것은 TM이 93,161 문장이었던 반면, 전처리에 의해 91,392 문장으로 1,769 문장이 중복되어 줄어들었다는 것이다. 중복된 문장이 줄어들었음에도 불구하고 200 문장의 번역률은 37 문장이 일치하여 5.34%가 오른 79.78%가 된 것을 알 수 있다. 다음의 예는 전처리를 적용하기 전과 후의 개선된 구체적인 실례이다.

(24) 전처리가 적용된 예

a. I'm sorry.

[적용 전] 나는 미안하다. (점수: 3.5)

[적용 후] 죄송합니다. (점수: 4.0)

b. You'll find it on your right.
[적용 전] 당신은 오른편에 그것을 발견할 것이다. (점수: 4.0)
[적용 후] 오른쪽에서 찾으실 수 있어요. (점수: 4.0)

또한 고유명사, 숫자, 기본 명사구를 변수로 치환함으로써 91,392 문장이었던 전처리 결과가 58,744 문장이 되어 32,648 문장이 중복되어 줄어든 것을 알 수 있다. 번역률도 82.03%가 되어 전처리의 번역률보다 2.25%가 향상된 것을 알 수 있다. 다음의 예는 고유명사, 숫자, 기본 명사구를 적용하기 전과 후의 개선된 구체적인 실례이다.

(25) 고유명사를 변수로 치환하여 적용된 예
a. One ticket to London, please.
[적용 전] 우리는 부디 London에 딱지를 붙인다. (점수: 1.5)
[적용 후] 런던 행 표 한 장 주세요. (점수: 4.0) (PRN1: London)
b. You might try the Park Hotel or the Morrison Motel.
[적용 전] 당신은 Park Hotel 또는 Morrison Motel을 시도할 것이다. (점수: 3.5)
[적용 후] Park 호텔이나 Morrison 모텔을 이용하실 수 있을 겁니다. (점수: 4.0) (PRN1: Park Hotel, PRN2: Morrison Motel)

(26) 숫자를 변수로 치환하여 적용된 예
a. 64 pounds, please.
[적용 전] 64는 부디 세게 두드린다. (점수: 1.5)
[적용 후] 64 파운드 주세요. (점수: 4.0) (NUM1: 64)
b. Walk over one block.
[적용 전] 하나의 블록을 무시하시오. (점수: 2.3)
[적용 후] 1 블록 걸어가세요. (점수: 4.0) (NUM1: one)

(27) 기본 명사구를 변수로 치환하여 적용된 예
a. I'm here on business.
[적용 전] 나는 사업에 여기에서 있다. (점수: 2.5)
[적용 후] 나는 사업에 여기에 왔습니다. (점수: 4.0)(BNP1:

business)

b. No, the one on the left.
[적용 전] 아니오, 왼쪽의 그 하나. (점수: 2.3)
[적용 후] 아니오, 왼쪽 위의 그 하나요. (점수: 3.0)(BNP1: the_one, BNP2: the left)

숙어 기반에 대한 TM 중복은 시스템이 현재 개발 중인 관계로 시스템에 의한 확인을 하지 못하였다. 따라서 200문장의 평가문에 대해서는 수작업으로 숙어 기반 방법이 적용되리라 생각되는 예문에 대해 시뮬레이션을 실시하였다. 그 결과 3 문장이 일치하면서 0.37%가 오른 82.40%가 된 것을 알 수 있었다. 다음의 예는 숙어 기반 방법을 적용하기 전과 후의 개선된 구체적인 실례이다.

(28) 숙어 기반 TM이 적용된 예

a. Thank you for choosing San Felice Hotel and have a nice day.
[적용 전] San Felice Hotel을 선택하여 주셔서 당신에게 감사하고 날씨가 좋은 날을 가지고 있으시오. (점수: 2.8)
[적용 후] San Felice 호텔을 선택하여 주셔서 당신에게 감사하고 좋은 시간 보내세요.(점수: 3.5) (VP1: choosing San Felice Hotel)

b. Would you like something else while you´re waiting?
[적용 전] 당신이 기다리고 있는 동안 어떤 다른 것을 좋아할 것인가? (점수: 2.5)
[적용 후] 당신이 기다리고 있는 동안 그 밖에 무엇을 드릴까요? (점수: 4.0)(SBAR1: while you're waiting)

5. 결론

본 논문은 두 가지를 목표로 하였다. 하나는 문자열 위주의 번역 메모리 (string-based translation memory, 본 논문에서는 TM으로 명명하였음)가 갖는 낮은 적용률을 높이기 위해 언어학적 구조가 부여된 구조화된 번역 메모리 (structured translation memory, 본 논문에서는 TM^{+}로 명명하였음)를 개발하는

것과, 다른 하나는 TM+와 영한 자동 번역 시스템을 연동하여 번역 품질이 개선되는 것을 확인하는 것이었다. 기존의 TM이 활용형 형태의 대역 말뭉치였다면, TM+는 TM과 더불어 언어학 구조를 변수로 도입한 형태를 가진다. 언어학 구조를 변수로 가지는 형태는 단계별로 형성되며 그 단계는 다음과 같이 이루어진다. 1) 전처리 2) 문두 부사 상당 어구 제거 및 확장 3) 고유 명사 청킹 및 PRN 치환 4) 숫자 청킹 및 NUM 치환 5) 기본 명사구 청킹 및 BNP 치환 6) 숙어 청킹 및 나머지 부분 변수 치환.

TM+의 효과를 검증하기 위해 200 문장의 관광용 영한 대화체 문장을 대상으로 TM과 비교 실험을 하여 보았다. TM을 적용한 결과 전문 번역가에 평가된 번역률은 74.44%였던 반면, TM+를 적용한 결과 번역률은 82.40%였다. TM+가 TM보다 7.96% 번역률을 향상시켰는데, 단계별로 보면 전처리 단계에 의해 5.34%, 고유명사/숫자/기본명사구 단계에 의해 2.25%, 숙어 청킹 단계에 의해 0.37%가 각각 번역률을 개선시켰다는 것을 알 수 있었다.

향후에 본 논문을 개선하는 계획에는 TM+의 단계를 더욱 세분화하는 것과 Elita et.al.(2006, 49)가 제안한 것과 같은 의미 정보를 고유 명사나 숫자 표현에 도입하여 더욱 정교한 번역을 이루도록 하는 것이다. 또한 TM+의 원문과 번역문 간의 정렬 기술을 더욱 개선하여 TM+를 반자동에서 자동으로 점진적으로 구축하도록 할 계획이다. 번역의 방향성과 관련해서 본 논문에서는 영어에서 한국어로의 방향이 제시되었는데 향후에는 한국어에서 영어로 가는 방향성에 대해 제시하고자 한다. TM+는 Trados와 같은 번역가 지원도구에서 직접 활용 가능하며 향후에 전문 번역가들이 이전에 번역해 두었던 번역문을 기존의 번역가 지원도구보다 더 많이 재활용할 수 있기를 바란다.

참고문헌

박주형, 이창우, 강명주. 2001.「자동 번역과 CAT의 현황과 전망」.『정보과학회지』제19권 제10호 특집 언어정보산업. 19-26.

Carl, Michael and Hansen, Silvia. 1999. “Linking Translation Memories with Example-Based Machine Translation”. *Proceedings of Machine*

Translation Summit VII – MT in the Great Translation Era. 617-624.

Elita, Natalia and Gavrila, Monica. 2006. "Enhancing Translation Memories with Semantic Knowldege". *Proceedings of the first Central European Student Conference in Linguistics.* Budapest. 49-54.

Hodasz, Gabor, Groebler, Tamas, and Kis, Balazs. 2004. "Translation Memory as a Robust Example-based Translation System". *Proceedings of 9th EAMT Workshop "Broadening Horizons of Machine Translation and its Applications"*. Malta, 82-89.

Hutchins, John. 2005. "Towards a definition of example-based machine translation". *MT Summit X, Proceedings of Workshop on Example-based Machine Translation*. Phuket, Thailand. 63-70.

Lagoudaki, Elina. 2006. "Translation Memories Survey 2006: Users' perceptions around TM use". *Proceedings of the Twenty-eighth International Conference on Translating and the Computer*. London. 1-29.

Planas, Emmanuel and Furuse, Osamu. 1999. "Formalizing Translation Memories". *Proceedings of Machine Translation Summit VII – MT in the Great Translation Era.* 331-339.

Rapp, Reinhard. 2002. "A Part-of-Speech-Based Search Algorithm for Translation Memories". *Proceedings of third International Conference on Language Resources and Evaluation*. Spain. 466-472.

Schäler, Reinhard. 2001. "Beyond Translation Memories". *Proceedings of Machine Translation Summit VIII – Workshop on Example-Based Machine Translation.* 49-55.

Simard, Michel and Langlais, Philippe. 2001. "Sub-sentential Exploitation of Translation Memories". *Proceedings of Machine Translation Summit VIII.* 335-339.

Vogel, S. and Ney, H. 2000. "Construction of a Hierarchical Translation Memory". *Proceedings of the 18th International Conference on Computational Linguistics*. Saarbruecken, Germany. 1131-1135.

[Abstract]

Study on English-Korean Structured Translation Memory

Choi, Sung-Kwon and Kim, Young-Kil
(Electronics and Telecommunications Research Institute)

This paper aims at developing a structured translation memory TM^+ which can resolve the coverage problem of an string-based translation memory TM and enhance a translation quality of English-Korean machine translation system.

The existing TM is basically a type of bilingual corpus with full-form words, while TM^+ including TM consists of different translation memories made by the following steps: 1) pre-processing, 2) deletion of sentence-initial adverbs and expansion, 3) chunking of proper noun and substitution of it by a variable PRN, 4) chunking of numeral expression and substitution of it by a variable NUM, 5) chunking of base noun phrase and substitution of it by a variable BNP, and 6) chunking of idioms and substitution of the remainders by their corresponding variables.

We evaluated 200 test sentences to compare the translation accuracy of machine translation system with TM with one of machine translation system with TM^+. The experimental result shows that while the translation accuracy of machine translation system with TM is 74.44%, the translation accuracy of TM^+ amounts 82.40%. From this result we could know that TM^+ rose a translation quality of 7.96%.

In the near future we have plans to segment the steps of TM^+ in more detail, introduce the semantic information for numeral expression, and develop the alignment technology between source segments and their target equivalents. We hope that TM^+ be applicable to the computer-aided translation tools like TRADOS and allow the professional translators to translate easier by using it.

▸Key Words: Translation Memory, Structured Translation Memory, Machine Translation, English-Korean Machine Translation

최승권
한국전자통신연구원 책임연구원
choisk@etri.re.kr
관심분야: 자동번역, 언어처리

김영길
한국전자통신연구원 책임연구원
kimyk@etri.re.kr
관심분야: 자동번역, 언어처리

논문투고일: 2009년 07월 15일
심사완료일: 2009년 08월 20일
게재확정일: 2009년 09월 05일