한국어 번역 메모리 시스템의 실현성 분석 및 설계

류철 [0] 노윤형 이기영 최승권 박상규
연 어 공 학 연 구 부
한 국 전 자 통 신 연 구 원
{ryuch, yhnoh, kylee, choisk, parksk}@etri.re.kr

Feasibility Test and Design of Korean Translation Memory System

Ryu, Cheol[0] Yoon-Hyung Roh Ki-Young Lee Sung-Kwon Choi Park. Sangkyu
Department of Linguistic Engineering
Electronics and Telecommunications Research Institute

요 약

번역 메모리(Translation Memory) 시스템이란 기존에 번역된 결과를 담고 있는 대용량의 번역 메모리에서 사용자가 제시한 입력문과 가장 유사한 문장을 검색한 후, 유사도 순으로 결과를 제시하여, 이후의 번역 작업을 보다 효율적으로 할 수 있도록 도와주는 시스템을 말한다. 이는 기계 번역 시스템과 비교해 볼때, 보다 실현 가능성이 높은 자연어 처리의 응용 분야라고 할 수 있다. 일반적으로 번역 메모리 시스템에서 핵심이 되는 요소는 번역 메모리의 구성과 유사성 척도에 대한 정의라고 할 수 있다. 국외의 경우, 이미 많은 상용 시스템들이 개발되어, 번역 작업의 시간 및 비용을 줄이는데 많은 도움을 주고 있지만, 국내의 경우, 한국어 번역 메모리의 구성 및 한국어 문장간 유사성 척도 등에 대한 연구가 미흡한 실정이다. 따라서 본 논문에서는 한국어를 대상으로 번역 메모리의 효율적인 구성 방법 및 문장간 유사성 척도에 대한 정의를 내리며, 한국어를 대상으로한 번역 메모리 시스템에 대한 실현 가능성을 논한다.

1. 서론

번역 메모리(Translation Memory) 시스템이란 원시 언어 문장과 이미 번역된 해당 대역문으로 구성된 대규모 데이터베이스를 기반으로 하여 사용자가 번역하고자 하는 입력 문장이 주어졌을 때, 이와 가장 유사한 문장을 데이터베이스에서 찾아주는 시스템을 말한다. 번역 메모리 시스템을 사용할 경우, 사용자는 번역 작업을 보다 빠르고, 보다 효율적으로 번역 작업을 수행할 수 있다. 이러한 번역 메모리 시스템은 기계 번역 시스템과 비교해 볼 때 보다 실현 가능성이 높고, 상품 가능성이 높은 분야라고 할 수 있다.

일반적으로 번역 메모리란 이전에 이미 번역한 문장과 해당 대역문의 쌍으로 구성된 데이터베이스를 말한다. 번역 메모리 시스템을 설계할 때, 고려해야 할 중요한 점은, 번역 메모리의 구성(format)과 문장간 유사도를 정의하는 유사성의 척도(similarity measure)이다. 초기 번역 메모리 시스템은 표층 어휘 레벨에서의 스트링 비교가 주를 이루었지만, 현재는 원시 문장에 대해 형태소 분석 단계 이상의 분석된 결과를 번역 메모리로 구성하여 사용하고 있다. 이 경우, 분석의 깊이가 깊을수록 분석에 의한 오류 및 경제적 비용을 감수해야 하며, 분석의 깊이가 얕을수록 단순 문자열 비교에 가까워져서 제한된 크기의 번역 메모리를 사용해서는 만족스러운 매칭률을 얻기 힘들다.

국외의 경우, 번역 메모리 시스템에 대한 연구가 이미 활발하게 이루어져서 자국의 언어를 원시 언어로 하는 상품화된 번역 메모리 시스템이 이미 상용화되어 사용되고 있다. 하지만, 이러한 외국어를 원시 언어로 하는 번역 메모리 시스템을 한국어에 그대로 적용할 경우, 한국어의 특성으로 인해 만족할 만한 성능을 기대하기 어렵다. 따라서 본 논문에서는 한국어의 특성을 분석하여 한국어에 알맞은

번역 메모리의 구성 방식을 제안하며, 동시에 한국어의 특성을 고려한 문장간 유사성 척도에 대해 제안한다. 또한 제안된 방법에 대한 실현 가능성을 실험을 통하여 분석한다.

2. 관련 동향

국내에서 한국어를 입력으로 하는 번역 메모리에 관한 연구는 본격적으로 이루어 지고 있지 않으나, 그 필요성에 대한 요구가 커지고 있다. 국내에서는 예제 기반 번역 시스템에 대한 연구가 번역 메모리의 연구와 유사하여 예제 기반 번역 시스템의 개발에 이용되었던 한국어 분석 기술과 유사 문장 검색 기술이 활용될 수 있을 것으로 보인다. 국외에서는 최근 발표된 논문인 Formalizing Translation Memory[1]에서는 번역 메모리에 대한 논의를 TELA 구조로 일반화 시키고 있고, Timothy Baldwin의 논문[3]에서는 일본어에서 의 문자 레벨에서의 번역 메모리 구현과 단어 레벨에서의 번역 메모리 구현이 큰 차이가 없음을 보여주었다.
본 논문은 Emmanuel Planas의 박사학위 논문[4]의 계층별 접근법에 착안하여 한국어에 같은 방법을 적용하여 보았다.

3. 실험 목표

본 실험에서는 한국어로 쓰여진 정보통신 분야의 기술 매뉴얼을 대상으로 번역 메모리의 활용가능성을 예측한다. 한국어는 영어와 다른 특성을 가지고 있고, 또한 기술 매뉴얼도 일반적인 분야와 다른 특성을 가질 것으로 생각할 수 있다. 실험은 번역 메모리 시스템을 계층적으로 설계 했을 때의 효과를 검증한다.
번역 메모리를 계층적으로 설계할 때에는 주로 어휘 계층, 원형 복원 형태소 계층, 품사 계층, 등등을 생각할 수 있는데, 본 한국어의 특성에 맞는 계층으로 어휘 계층, 원형 복원 형태소 계층, 지배 명사 계층, 명사 계층의 4개의 계층에 초점을 맞추었다. 품사 계층은 문맥의 자유도가 높은 한국어의 특징에 적합하지 않아서 고려 대상에서 제외되었다.
또한 번역 메모리가 번역 작업에 기여할 수 있는 실제적인 영역이 무엇인지 확인하는 실험을 수행하였다. 일반적으로 번역 메모리는 범용적으로 사용할 수 없고, 특정한 분야나 목적 별로 번역 메모리를 구축하여 사용하는데, 한국어로 만들어진 정보통신 분야의 경우에서도 같은 특징을 가지는지 확인하였다.

4. 실험 방법

본 실험에서 대상 문장은 정보통신분야 기술 매뉴얼을 가공하여 사용하였다. 매뉴얼은 제목들과 표, 그림 등이 포함되어 있으나, 이들은 실제로 번역의 어려움이 없으나 문장의 형태를 가진 설명은 번역 메모리의 도움이 더 요청된다. 매뉴얼 텍스트 전체를 형태소 분석하여 나온 결과를 이용하여 이중에 설명에 해당하는 부분만을 추출하여 실험에 사용하였다. 매뉴얼 텍스트는 약 2만 8천 문장이고, 평균 문장의 길이는 9.6 어절이다.
실험은 각 계층별로 매칭률을 계산한다. 매칭률이란 임의의 문장을 선택했을 때 전체 구축된 번역 메모리에서 같은 문장에서 발견할 확률을 나타내는 것으로 매칭률이 높을수록 번역 메모리를 구축했을 때 생산성이 높아지는 효과를 기대할 수 있다.
각 계층은 다음과 같이 구성한다. 어휘 계층은 원문을 형태소로 분해하여 사용하고, 원형 복원 형태소 계층은 원문을 형태소로 분해하고 복원 가능한 형태소는 복원하여 사용한다. 그것은 명사의 경우 복수 접미사나 동사의 경우 선어말어미 등을 제거함으로 이루어진다. 그리고, 지배 명사 계층은 명사구에서 지배 명사로 간주되는 가장 뒤의 명사를 제외한 명사를 제거하여 사용하고 명사 계층은 모든 명사구를 구문 심볼로 치환하여 사용한다. 이 과정을 위해 본 실험에서는 형태소 태깅 결과 중에서 조사, 어미, 용언만을 추출하여 사용하고, 지배 명사는 조사 바로 앞의 명사만을 추출하여 구한다.
매칭이란 문장 전체가 정확히 일치하는 것을 말하는 것으로 간단히 문자열을 비교하여 일치 여부를 판별하고 일치하는 문장별로 빈도수를 계산한다.
또한, 한국어 정보 통신 분야 매뉴얼에 대해서 일반적으로 알려진 번역 메모리의 특성이 적용될 수 있는지 확인하였다. 본 실험에서 사용된 매뉴얼은 정보 통신 분야의 세가지 제품에 대한 매뉴얼로 각 매뉴얼 별로 번역 메모리를 적용하는데 서로 다른

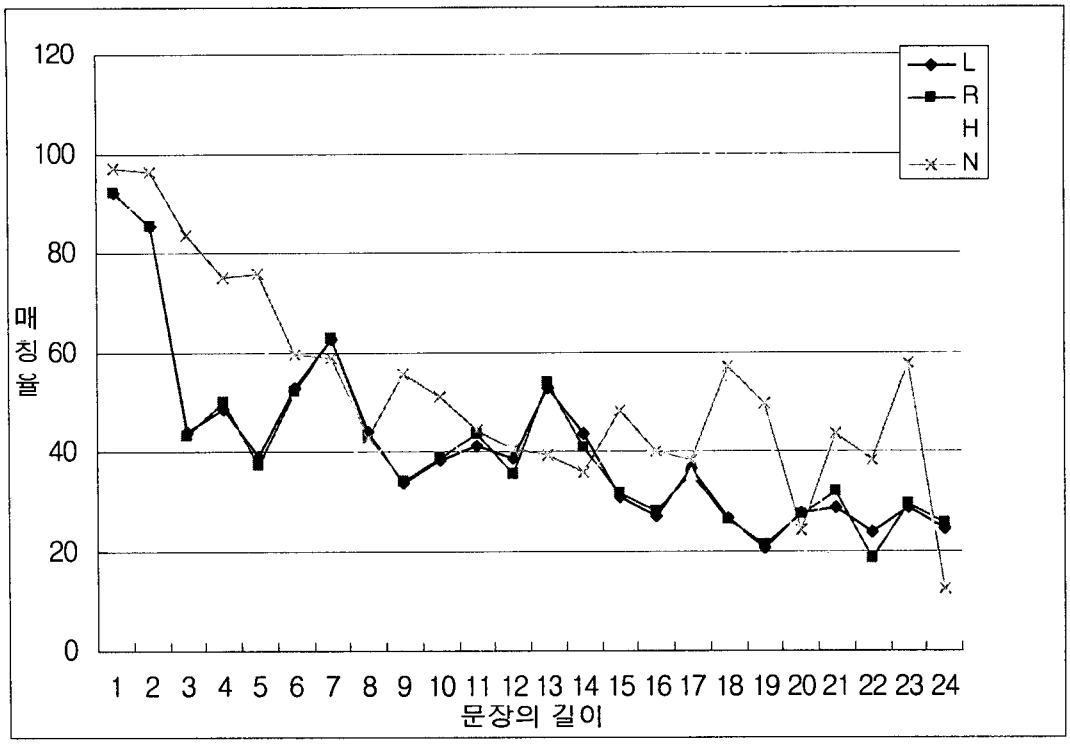

그림 1: 길이에 따른 문장 매칭률

특성이 있는지 확인하였다.

5. 실험 결과

[그림1]은 각 계층별로 문장의 길이에 따라 매칭율이 변화를 보여주고 있다. L, R, H, N은 각각 어휘 계층, 원형 복원 형태소 계층, 지배 명사 계층, 명사 계층을 나타낸다. 가로축은 문장의 길이를 나타내는 것으로, [그림1]에서 문장의 형태소 길이가 9인 경우에는 약 50%의 매칭율을 나타내고 있다. 이것은 형태소 길이가 9인 임의의 문장을 하나 골라 내었을 때 형태소 길이가 9인 번역 메모리에서 하나 이상의 이미 번역한 결과가 존재하여 이를 이용할 수 있는 확률이 0.5라는 것을 의미한다.

보통 영어 번역 메모리의 경우 어휘 계층에서 매칭한 후 동사나 명사 중 굴절이 일어난 단어들을 원형 형태소로 복원하여 매칭율을 높이고자 시도해 왔다. 그러나, [그림1]에서 발견한 한국어의 특성은 원형 형태소로 복원하여 번역 메모리 시스템을 만든다 하여도 영어에서 기대했던 것만큼의 효과를 거둘 수 없음을 말해준다. 즉, 원형 복원 형태소 계층과 어휘 계층이 거의 같은 매칭율을 가지고 있다는 것을 보여 주고 있다. 이것은 한국어가 굴절어가 아니기 때문에 나타나는 현상이다.

이에 비하여 지배 명사 계층은 평균 약 40%의 매칭율을 보임으로써 한국어의 복합 명사에서 맨 뒤의 지배명사를 이용하여 번역 메모리를 구현하면 매칭율을 크게 향상 시킬 수 있다는 것을 보여주고 있다. "한영번역시스템을 위한 문틀기반번역방식의 실현성 분석" [1]에서는 일반적인 텍스트에서 수행한 명사 계층의 매칭과 유사한 실험에서 매우 낮은 수치를 얻었지만, 본 논문에서는 정보통신 기술 매뉴얼이라는 특정한 분야를 선택해서 같은 실험을 했을 때에는 평균 50%의 매칭율을 보였다. 이것을 주목하여 보면 일반적인 텍스트에서는 같은 문장이 되풀이 될 가능성이 전혀 없지만, 기술 매뉴얼과

같은 문서에서는 같은 문장이 여러 번 사용되고 있음을 확인 할 수 있다. 또한 문장 길이에 따라서 매칭율이 크게 변화하지 않는 것은 긴 문장이라도 본 실험이 선택한 기술 매뉴얼과 같은 문서에서는 중복되어 사용되는 경우가 많기 때문으로 보인다. 이것은 번역 메모리를 잘 활용할 경우 긴 문장이라도 쉽게 기존 번역을 이용하여 빠른 시간 안에 번역할 수 있음을 알려준다. 따라서, 번역 메모리의 구성을 위해서는 대용량의 일반 영역 코퍼스보다는, 주제와 용도가 같은 문서에서 얻을 수 있는 번역 예가 더 효과적으로 사용될 수 있음을 알 수 있다.

[그림2]는 계층별 매칭율로써, 계층별 매칭율은 각 계층의 전체 문장 수에 대하여 그 계층에서 중복되는 문장 수의 비율을 의미한다. 즉, 하나의 문장에 대한 유사한 번역의 예가 존재할 수 있는 확률을 나타내는 것으로 볼 수 있다. 예를 들어 어휘 계층에서 매칭율이 0.4라는 것은 새로 번역할 문장과 유사한 문장이 번역 메모리에 들어있을 평균 확률이 0.4라는 것으로 40%의 번역을 위한 수고가 덜어질 수 있다. 그림에서 어휘계층, 지배 명사 계층, 명사 계층으로 일반화를 많이 시킬수록 매칭율이 높아 지는 것을 보여주고 있다.

[그림2]에서 어휘 계층과 원형 복원 형태소 계층은 거의 같은 매칭율을 보이고 있어 원형 복원 형태소 계층은 번역 메모리 시스템을 구현하는데 별 유용성을 주지 못하는 반면, 지배 명사 계층은 매칭율이 높고 지배 명사를 중심으로 명사구를 한정하므로 정확율이 높으므로 번역 메모리 시스템을 구현하는데 유용할 것으로 보인다. 이것은 한국어의 경우 중심어가 기능어와 연결되는 특징을 가지고 있는데, 이를 이용함으로써 지배 명사를 쉽고 정확하게 추정할 수 있기 때문에 지배 명사 계층을 이용한 번역 메모리는 한국어의 특성을 잘 고려한 실현성이 높은 시스템이 될 수 있다.

[그림3]은 명사 계층에서 매칭된 문장들의 예를 보인 것으로

NP 로/41 조회하/60 는/93 NP 의/42 NP 를/41 알/60 다/91

라는 구조에 매칭된 문장들을 볼 수 있다. 여기에서 슬래쉬 뒤의 숫자는 품사를 나타내고, NP는 명사가 올 수 있음을 나타낸다. 지배 명사 계층에서 이 구조에 매칭된 문장들은 다음과 같다.

명령어/12 로/41 조회하/60 는/93 GAN/11 의/42 정보/12 를/41 알/60 다/91

명령어/12 로/41 조회하/60 는/93 기지국/12 의/42 정보/12 를/41 알/60 다/91

명령어/12 로/41 조회하/60 는/93 링크/12 의/42 정보/12 를/41 알/60 다/91

명령어/12 로/41 조회하/60 는/93 제어국/12 의/42 정보/12 를/41 알/60 다/91

즉, 첫번째 명사구의 지배어로 "명령어"가, 두번째 명사구의 지배어로 "GAN", "기지국", "링크", "제어국" 등이, 세번째 명사구의 지배어로 "정보"가 오고 있음을 볼 수 있다. 그리고 각각의 지배 명사 계층에 속하는 어휘 계층에 문장들은 그림에서 보는 것과 같다. 따라서 다시 명령어를 하는 수식어로 "CONFIGURATION"-"Dis_SysConf"-"DIS-GEP-CONF" 등이, 정보를 한정하는 수식어로 "GEP", "GMP", "GOP" 등이 올 수 있음을 알 수 있다. [그림3]에서 보면 실제 동일한 구조의 문장들이 명사구만 바뀌어 쓰이는 것을 볼 수 있고, 기존의 대역문에서 필요한 부분만 치환함으로써 원하는 문장의 번역문을 얻을 수 있다.

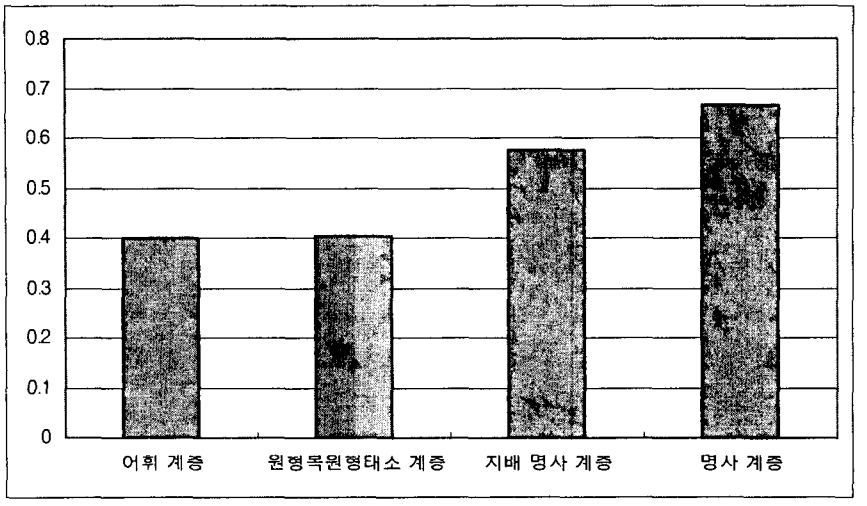

그림 2: 계층별 매칭율

```
NP 로/41 조회하/60 는/93 NP 의/42 NP 를/41 알/60 다/91

명령어/12 로/41 조회하/60 는/93 GAN/11 의/42 정보/12 를/41 알/60 다/91
        "CONFIGURATION"-"Dis_SysConf"-"DIS-GEP-CONF" 명령어로 조회하고자 하는 GAN의 GEP 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-GMP-CONF" 명령어로 조회하고자 하는 GAN의 GMP 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-GOP-CONF" 명령어로 조회하고자 하는 GAN의 GOP 정보를 알 수 있다.
명령어/12 로/41 조회하/60 는/93 기지국/12 의/42 정보/12 를/41 알/60 다/91
        "CONFIGURATION"-"Chg_BssPara"-"CHG-BTU-PARA" 명령어로 조회하고자 하는 기지국의 BTU 정보를 알 수 있다.
        "CONFIGURATION"-"Chg_BssPara"-"CHG-FA-PARA" 명령어로 조회하고자 하는 기지국의 BTU 정보를 알 수 있다.
        "Configuration"-"Chg_BssPara"- "CHG-SECT-PARA" 명령어로 운영자가 조회하고자 하는 해당 기지국의
        Sector Parameter 정보를 알아본다.
명령어/12 로/41 조회하/60 는/93 링크/12 의/42 정보/12 를/41 알/60 다/91
        "CONFIGURATION"-"Dis_SysConf"-"DIS-BLNK-CONF" 명령어로 조회하고자 하는 링크의 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-GLNK-CONF" 명령어로 조회하고자 하는 링크의 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-LLNK-CONF" 명령어로 조회하고자 하는 링크의 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-MLNK-CONF" 명령어로 조회하고자 하는 링크의 정보를 알 수 있다.
명령어/12 로/41 조회하/60 는/93 제어국/12 의/42 정보/12 를/41 알/60 다/91
        "CONFIGURATION"-"Dis_SysConf"-"DIS-AEP-CONF" 명령어로 조회하고자 하는 제어국의 AEP 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-AMP-CONF" 명령어로 조회하고자 하는 제어국의 AMP 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-AOP-CONF" 명령어로 조회하고자 하는 제어국의 AOP 정보를 알 수 있다.
        "CONFIGURATION"-"Dis_SysConf"-"DIS-ATP-CONF" 명령어로 조회하고자 하는 제어국의 ATP 정보를 알 수 있다.
```

그림 3: 명사 계층에서의 매칭 예

일반적으로 알려진 번역 메모리의 특성인 한국어로 된 정보 통신 분야 매뉴얼에서도 적용되는지 확인하기 위하여 준비된 세 제품의 매뉴얼에 각각에 대하여 매칭률을 얻었다. [그림4]를 보면 매뉴얼 전체에 대한 매칭률과 큰 차이가 나지 않는 것을 알 수 있는데, 이것은 각 제품의 매뉴얼이 공유하는 문장의 수가 많지 않고, 되풀이 되는 문장들은 같은 제품의 매뉴얼에 존재한다는 것을 추측할 수 있다.

[그림5]는 문서간 매칭률과 문서내 매칭률을 비교한 것으로 매뉴얼 A의 문서내 매칭률과 A를 제외한 나머지 매뉴얼과 A의 매칭률을 이 매우 큰 차이

를 나타내고 있음을 알 수 있다. F의 경우도 마찬가지인 것을 알 수 있는데, [그림4]가 문서내 매칭률이 매우 높음을 나타내는데 비하여 [그림5]는 문서간 매칭률은 매우 낮음을 나타낸다. 즉, 번역 메모리를 구성하여 번역 작업에 이용할 때에 아무리 같은 분야라도 기술하는 제품이 다른 경우에는 활용도가 매우 낮고, 같은 제품을 기술하거나 또는 개정판을 낼 때에는 40 ~ 50 %의 높은 매칭률을 얻을 수 있을 것을 알 수 있다.

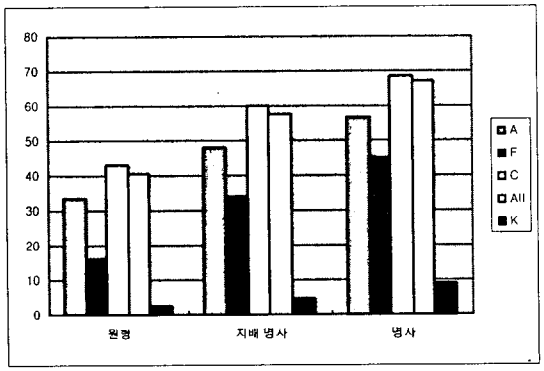

그림 4: 문서내 매칭률

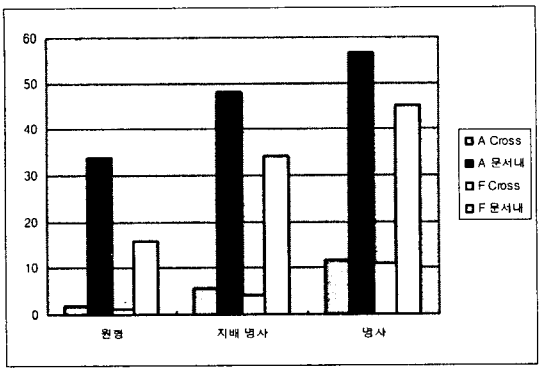

그림 5: 문서간 매칭률과 문서내 매칭률 비교

6. 유사성 척도

앞에서 살펴보았듯이, 실제 매뉴얼에서 사용하는 문장들에 있어 주로 명사구가 치환된 문장 형태들이 많이 존재함을 알 수 있다. 따라서, 명사구에서 중심어를 제외한 수식어를 제거함으로써 높은 매칭율과 정확성 손실을 최소화할 수 있고, 이를 잘 이용할 수 있도록 번역 메모리 형태와 유사성을 정의해야 한다. 이러한 번역 메모리 및 유사성을 위해 다음과 같은 사항을 고려하고자 한다. 먼저 유사성 판정을 위해 사용하는 기본 알고리즘은 편집 거리(Edit Distance)이다. 편집 거리란 한 문장을 편집하여 다른 문장으로 수정하기 위해 드는 편집 연산의 비용으로 계산되며, 문장 어순에 민감한 척도로써 유사 문장 판정을 위해 상당히 좋은 성능을 나타낸다. 본 논문에서는 편집 거리를 이용하여 각각의 형태소를 매칭할 때, 각 형태소의 품사 종류에 따라 가중치를 주어 유사도를 판정하고자 한다. 먼저, 명사 계층을 구성하는 요소로써 조사, 어미, 용언에 가장 큰 가중치를 두하고, 그 다음 지배 명사 계층을 구성하는 지배 명사에 그 다음 가중치를 부여하고, 그리고 나머지 형태소들에 가장 낮은 가중치를 부여한다. 이렇게 함으로써 문장의 유사성을 판정할 때, 문장 구조에 큰 영향을 끼치는 요소에 더 큰 가중치를 부여함으로써 좀 더 의미 있는 유사 문장을 추출할 수 있게 된다. 예를 들어 두 문장에서 하나의 명사가 다를 경우 지배 명사로 쓰인 명사보다는 수식어로 쓰인 명사가 다를 경우 더 유사한 문장으로 판정할 수 있는 것이다.

7. 결론

본 논문에서는 한국어 번역 메모리 시스템의 실현성 및 설계를 위해 기술 매뉴얼에 대해 어휘 계층, 원형 복원 형태소 계층, 지배 명사 계층, 명사 계층에 대해 각 계층별 문장 매칭율에 대해 실험하였다. 그 결과, 한국어에서는 원형 복원이 별 유용성을 제공하지 못하고, 지배 명사를 추출하여 이루어지는 지배 명사 계층에서 큰 매칭율의 향상이 있을 뿐아니라, 분석의 용이성 및 매칭의 정확성으로 유용성이 있음을 제시했다. 그리고, 일반 영역과 달리 기술 매뉴얼 문서에서 굉장히 높은 어휘 계층에서의 매칭율을 보였는데, 이는 번역 메모리 시스템이 일반 영역보다는 기술 매뉴얼 분야와 같은 특정 영역에서 잘 적용될 수 있고, 명사구까지의 치환을 고려할 때, 한국어에서 충분한 유용성을 갖고 있음을 보여 준다. 그리고 앞의 계층 구조의 특성을 고려하여 형태소의 기능에 따라 가중치를 부여하는 유사도 판정 방법을 제시하였다. 문서내 매칭률과 문서간 매칭률을 측정해 본 결과 문서내 매칭률이 매우 높고, 같은 분야라도 문서간 매칭률이 다른

일반 문서와 가지는 매칭률과 같이 매우 낮은 것을 확인하여 번역 메모리가 적용될 수 있는 번역 작업의 한계를 알 수 있었다.
앞으로 남은 과제로는 무엇보다 계층구조의 구성이 형태소 분석 및 태깅 결과를 통해 이루어지기 때문에 높은 형태소 분석기의 성능이 요구되고 이에 대한 형태소 분석기의 성능향상이 요구된다. 또한 구체적인 설계를 통한 유사도 판정을 위한 각 형태소별 가중치에 대한 실험적인 연구가 필요하다고 생각된다.

8. 참고 문헌

[1]. 김영길, 서영애, 서광준, 최승권, "한영번역 시스템을 위한 문틀기반번역방식의 실현성 분석", 2000년 한국정보처리학회 추계 학술발표논문집
[2]. Emmanuel Planas, Osamu Furuse, "Formalizing Translation Memories"
[3]. Timothy Baldwin, Hozumi Tanaka, "The Effect of Word Order and Segmentation on Translation Retrieval Performance"
[4]. Emmanuel Planas, "TELA: Structures and Algorithms for Memory-based Machine Translation"