도메인 특화 방법에 의한 영한 특허 자동 번역 시스템의 구축

(Construction of English-Korean Automatic Translation System for Patent Documents Based on Domain Customizing Method)

최 승 권 † 권 오 욱 †† 이 기 영 †† 노 윤 형 †† 박 상 규 ††
(Sung-Kwon Choi) (Oh-Woog Kwon) (Ki-Young Lee) (Yoon-Hyung Roh) (Sang-Kyu Park)

요 약 본 논문은 웹과 같은 일반적인 도메인의 영한 자동 번역기를 특정 도메인으로 특화하는 방법에 의해 구축된 영한 특허 자동번역 시스템을 기술하는 것을 목표로 한다. 특정 도메인으로서의 특허 자동번역기를 위한 특화 방법은 다음과 같은 단계로 이루어진다: 1) 대용량 특허 문서의 수집 및 언어학적 특성 분석, 2) 전문용어 추출 및 대역어 구축, 3) 기보유한 용어의 대역어 특화, 4) 특허 고유의 번역 패턴 추출 및 구축, 5) 언어학적 특성 분석에 따른 기보유 번역 엔진 모듈의 특화 및 개선, 6) 특화된 번역 지식 및 번역 엔진 모듈에 따른 번역률 평가. 이와 같은 특화 절차에 따른 특허 영한 자동 번역기는 특허 전문번역가의 평가에 의해 전분야 평균 81.03%의 번역률을 내었으며, 분야별로는 기계(80.54%), 전기전자(81.58%), 화학일반(79.92%), 의료위생(80.79%), 컴퓨터(82.29%)의 성능을 보였으며 계속 개선 중에 있다.

키워드 : 특화 방법, 도메인 특화, 자동 번역, 특허 번역, 영한 번역, 특허 자동 번역

Abstract This paper describes an English-to-Korean automatic translation system for patent documents which is constructed by a method customizing from a general domain to a specific domain. The customizing method consists of following steps: 1) linguistically studying about characteristics of patent documents, 2) extracting unknown words from large patent documents and terminologically constructing, 3) customizing the target language words of existing terms, 4) extracting and constructing patent translation patterns peculiar to patent documents, 5) customizing existing translation engine modules according to linguistic study about characteristics of patent documents, 6) evaluation of automatic translation results. The English-to-Korean patent machine translation system implemented by these customization steps shows a translation accuracy of 81.03% and is improving.

Key words : Customizing method, Domain customization, Automatic translation, Patent translation, English-Korean translation, Machine translation for patent documents

1. 서 론

영한 자동 번역기가 국내에서 연구 개발된 지도 벌써 10여 년이 지나고 있다. 그 동안 영한 자동 번역기의 응용은 텍스트로부터 웹, 방송자막, 모바일 등으로 확대되었으며, 영한 자동 번역기의 번역률은 웹과 같은 일반 도메인의 문어체 텍스트를 대상으로 현재 70% 정도의 성능을 보이고 있다. 그러나 영한 자동번역기의 번역률은 70%대를 정점으로 더 이상 크게 개선되고 있지 못한 실정이다.

외국의 경우에도 언어 유형이 상이한 이종 언어간의 자동번역 성능에 있어서 국내의 경우와 유사하게 일반 도메인을 대상으로 한 자동번역 결과가 일정 번역율을 내고서는 계속 답보상태에 있다.

이에 따라 국외에서는 90년대 초반부터 자동 번역기의 번역율을 향상시키기 위해 자동 번역기의 대상을 일반 도메인에서 제한된 도메인으로 특화하는 시도를 하였으며, 비교적 정확한 대역어 선택을 함으로써 일반 도메인보다도 상당히 유용한 번역 결과를 생성해 낼 수 있었다.

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

† 정 회 원 : 한국전자통신연구원 언어처리연구팀 연구원
choisk@etri.re.kr
†† 비 회 원 : 한국전자통신연구원 언어처리연구팀 연구원
ohwoog@etri.re.kr
leeky@etri.re.kr
yhroh@etri.re.kr
parksk@etri.re.kr

논문접수 : 2006년 8월 13일
심사완료 : 2006년 8월 31일

국내에서는 일반 도메인을 대상으로 한 영한 자동번역기가 앞서 언급한 바와 같은 다양한 응용영역에 응용되고 있음에도 불구하고 아직도 사용자들이 현재의 영한 번역기의 번역품질에 만족하고 있지 못한 실정이다[1,2].

사용자들이 웹과 같은 일반 도메인을 대상으로 한 영한 자동 번역기를 만족스러워하지 않는 원인을 기술적 측면에서 요약하면 다음과 같을 것이다:

- 빈번하게 출현하는 대역어 선택 모호성: 번역 도메인을 제한하지 않음으로써 다양한 실세계 지식을 고려해야 하는 대역어 선택 모호성이 발생하며 사전에 등록해야 할 단어의 수가 무제한으로 많아짐.
- 편향성이 없는 형태소 및 구조적 모호성: 도메인 특성에 따른 품사나 구조의 편향성이 반영되지 못함으로써, 태깅과 파싱에서 품사 모호성 및 구조적 모호성이 발생하여 번역품질이 낮아짐.
- 장문 번역의 오류: 40단어 이상으로 구성되는 장문은 문장의 복잡도로 인해 구조적 모호성이 발생하여 잘못된 번역을 야기시킴.
- 번역 패턴의 낮은 적용률: 일반 도메인에서는 구축된 번역패턴의 적용이 낮음으로써, 자연스러운 번역을 제공하지 못함.

이러한 기술적 문제점을 해결하기 위해 본 논문에서는 일반 도메인을 대상으로 구현되었던 기존 영한 자동번역 시스템을 도메인 특화 방법에 의해 구현된 영한 특허 자동 번역 시스템에 대해 기술하고자 한다.

본 논문에서 기술되는 영한 특허 자동번역 시스템은 정보통신부가 2005년부터 2006년도까지 개발 지원한 선도기반 기술 과제의 일환으로 개발되었으며, 2005년도에는 전기전자 분야를 대상으로 한 전기전자분야 특허 영한 자동번역기가 개발되었으며, 2006년도에는 영어 특허 문서의 모든 분야를 망라하는 전분야 특허 영한 자동번역기가 개발되고 있다.

2. 특허 문서의 수집 및 언어학적 특성

일반 도메인의 시스템을 제한된 도메인의 시스템으로 특화하기 위해서는 우선적으로 제한된 도메인에 해당되는 원시 코퍼스를 대량으로 수집하여 특정 도메인에 고유한 언어학적 특성을 분석하는 것이 중요하다.

영어 특허 문서의 고유한 언어학적 특성을 찾기 위해 활용한 특허 문서는 1,001,419 건이다.[1] 이 특허 문서는 2001년부터 2005년까지 미국에서 공개 출원된 특허 문서이며 문장수로는 290,683,622에 달하며 1문서당 평균 문장수는 290 문장 정도이고, 1 문장당 평균 단어수는 28.12 단어였다.

영어 웹문서의 1문장당 평균 단어수는 [3]에 따르면 22단어이므로 웹 문장 보다 특허 문장의 평균단어수가 6단어 이상 많다는 것을 알 수 있다.

본 장에서는 상기에 언급된 영어 특허 문서 중에서 1/1,000에 해당하는 1,000건의 특허문서를 임의로 추출하여 언어학적 특성 분석을 실시하였다.

2.1 형태적 특성

일반 문서와 비교해서, 특허 문서에서는 기호 단어,[2] 수식, 약어, 고유명사와 하이픈 연결 단어[3]가 많이 나타난다. 품사 태깅 관점에서 볼 때 특허 도메인에서 특정 품사로 편향되는 특정 단어들이 있다. 다음은 특정 품사로 편향되는 예이다.

- 거의 모든 "-ed" 형태의 단어들이 과거형보다는 과거분사로 사용된다.
- 대명사 출현빈도가 일반 도메인보다 적으며, 특히 인칭대명사는 거의 나타나지 않는다.
- "said"가 특허 문서에서는 거의 모두 형용사로 사용되며 "die"는 동사보다는 명사로 대부분 사용된다.

특허 문서에는 일반 도메인에서와는 상당히 다른 품사 n-gram이 나타나는데, 대표적인 경우가 "NN(NNS) CD VBP(VBZ)" 형태이다. 일반 도메인에서는 수사 뒤에 명사가 오는 반면에, 특허에서는 "the selection *algorithm 212* processes 3 selection" 와 같이 청구항 번호를 지칭하기 위해 명사 뒤에 수사가 오는 특이한 형태의 n-gram이 빈번하게 나타난다.

2.2 구문적 특성

영어 특허문서의 구문적 특성을 정리하면 다음과 같다.

- 병렬 및 복잡한 수식에 의한 장문이 많음: 특허에서 40단어 이상의 장문이 전체 문장 대비 약 19.28%를 차지한다. 특히 특허에서 중요한 정보를 담고 있는 요약문이나 청구항의 경우 평균 약 35 단어이다.
- 특허 전형적인 상투적 구문 표현을 많이 사용함: 복합단어(present invention), Elaboration(A of Z which Y X), *Characterization*(characterized in that), Jepson-like Style(In X, Z which Y), Composition(composed of ~, ···, and ~) 등의 특허 고유의 상투적인 표현이 많음[4].
- 전치사구, 분사구의 부착에 있어서 편향성을 나타냄: for 전치사구 및 현재 분사구 수식의 경우 동사구 보

1) 수집된 특허문서는 ETRI에서의 연구 목적으로 특허청으로부터 MOU 체결에 의해 획득한 공개특허 문서이다.

2) 본 논문에서는 기호 단어란 프로그램, 수학/공학적 표기법에 사용되는 다양한 단어 및 기호를 말하며, 특히 "한글"로 번역되지 않는 단어를 의미한다.

3) 하이픈으로 연결된 단어, 예를 들어, *"color-modified", "green-sensing", "more-normal", "multi-processor"*, etc.

다는 명사구 부착의 편향성을 나타냄.

- 독립 분사 형태의 표현을 많이 사용함: NP VP-ing 형태의 문장이 앞에서 언급된 문장에 대해 더욱 상세히 설명하기 위해 자주 사용됨.
- 일반 도메인에서 많이 사용되는 시간부사구가 사용되지 않고, 도치, 생략 등의 특수한 구문이 비교적 적게 사용됨.

2.3 대역어의 특성

특허 문서에서 영어 표제어에 대한 대역어의 모호성은 두가지 형태로 나타난다. 첫째는 동일한 특허 분야에서 문맥에 따라 다르게 선택되어 번역되는 경우다. 이에 대한 예는 다음과 같다.

object@NOUN {목적, 대상}
condition@NOUN {조건, 상태}
treatment@NOUN {처리, 치료}
end@NOUN {목적, 단부}

두번째는 다른 특허 분야에서 다르게 선택되어 번역되는 경우다. 이에 대한 예는 다음과 같다.

표 1 특허 분야별 대역어 선택 모호성 어휘 예

어휘	의료위생 분야	기계 분야
body	몸	바디
face	얼굴	표면
operation	수술	작동
order	처방	정렬
case	증례	경우

2.4 특허 고유의 구문/문장 패턴

특허문서는 일반 문서와 달리 특허문서 고유의 구문 및 문장 패턴이 존재한다. 이러한 구문 및 문장 패턴은 각 Field별로 다음과 같이 여러 형태로 나타난다:[4)]

- Abstract: 특허 대상 기술을 소개하는 문장이 특허문서마다 등장한다.
  - The present invention relates to ~ => 본 발명은 ~에 관한 것이다.
  - A method for ~ comprises the step of, 1) …, 2) … => ~기 위한 방법은 1) …, 2) …의 단계로 구성된다.
- Background of the Invention: 도면을 참조하여 내용을 간략히 설명하는 문장이 등장한다.
  - As shown in FIG.~, => 도 ~에서 도시된 바와 같이,
- Summary of the Invention: 제안되는 발명의 목적을 설명하는 문장이 일반적으로 제일 처음 나온다.
  - Accordingly, it is an objective of the present invention to provide ~ : 따라서, ~을 제공하는 것이 본 발명의 목적이다.
- Claims: 주로 현재분사/동명사를 동반한 명사구가 대부분을 이룬다.
  - The method of ~ comprising the steps of: 다음의 단계들로 구성되는 ~는 방법

3. 특화 방법

자동번역 시스템의 번역율을 개선시키기 위해 국외에서는 90년대 초반부터 자동 번역기의 번역 대상 도메인을 제한된 도메인으로 특화하는 시도를 하기 시작했다.

이러한 시도의 한 예로써 다국어 자동번역 시스템인 SYSTRAN은 일반 도메인으로부터 제한된 도메인으로의 자동 번역 시스템을 특화하는 절차를 소개한 바 있다[5]. [5]에서 특화 프로세스는 다음과 같은 단계들로 구성된다: 1) 용어 추출(Term extraction) 단계: 미등록어, 기존 사전의 고빈도 어휘 추출, 하위범주화 어휘 패턴 등을 추출 2) 사전 특화(dictionary customization) 단계: 용어의 번역, 활용형 정보 코딩 등 3) 언어학적 특화(Linguistic customization) 단계: 동음이의어, 전치사구 부착, 접속사, 복합문 구성 등의 분석 4) 번역 결과와 관련된 평가(Testing and Evaluation) 단계.

상기의 특화 프로세스는 기존의 한영 자동번역 시스템을 특허 도메인에 특화하여 특허 한영 자동번역 시스템을 개발하는 데도 적용한 바 있다[6].

본 논문에서 기술되는 특화 방법과 기존의 특화 방법들과 비교할 때, 본 논문에서 기술되는 특화 방법은 기존의 특화 방법을 수용하면서도 다음과 같은 점에서 그 차별성 및 우수성을 가진다:

1) 용어 추출 단계에서, 기존의 특화 방법들이 미등록어 추출만을 다룬 반면, 본 논문의 특화 방법에서는 미등록어 추출 뿐만 아니라 기보유하고 있는 다른 전문용어의 활용도 특화 방법에 포함하고 있다.

2) 기보유하고 있는 번역사전 엔트리의 대역어를 특화함에 있어서, 기존의 특화 방법들에서는 기존 번역사전의 대역어 특화 방법이 구체적으로 어떻게 하는지에 대해 기술되어 있지 않으나, 본 논문에서는 기보유 엔트리의 대역어 특화를 두부분으로 나누어 기술한다. 하나는 고빈도에 따른 디폴트 대역어 선정 방법이며, 다른 하나는 특허 분야별 대역어 선정 방법에 관한 것이다.

4) 특허문서를 구성하는 Field들을 순서대로 기술하면 다음과 같다 : Title, Abstract, Technical Field, Background of the Invention, Summary of the Invention, Brief Description of the Drawings, Description of the preferred Embodiments, Claims. 이러한 Field들은 특허 발명자에 따라 일부가 생략되어 기술되기도 한다.

3) 특허 고유의 번역 패턴 추출 및 구축에 있어서, 기존의 특화 방법들에서는 영어 특허에 고유한 패턴에 대한 언급이 없으나, 본 논문에서는 영어 특허 문서 고유의 구문 및 문장 패턴에 대한 추출 및 구축에 대해 기술된다.

4) 기존 번역 엔진 모듈의 특화에 있어서, 기존의 특화 방법들은 영한 번역 엔진에 대한 특화 방법이 아닌 타 언어 엔진에 대한 특화 방법이 기술되어 있으나, 본 논문에서는 영한 번역 엔진을 특허 문서에 특화시키는 방법에 대해 기술한다.

본 절 이후부터는 기존의 특화 방법들을 포함하면서 상기에 기술된 추가적인 특화 방법에 대해 더 상세히 기술하고자 한다.

3.1 용어 수집, 추출 및 구축

일반 도메인에 적응되어 있는 자동번역 시스템을 특허 도메인에 특화된 특허 자동번역 시스템으로 전환시키기 위해 기존의 용어를 수집하고 특허 문서로부터 미등록 용어를 추출 및 구축하는 작업이 우선 이루어져야 한다. 이와 같은 용어 수집, 추출 및 구축은 다음과 같은 그림으로 기술될 수 있다:[5]

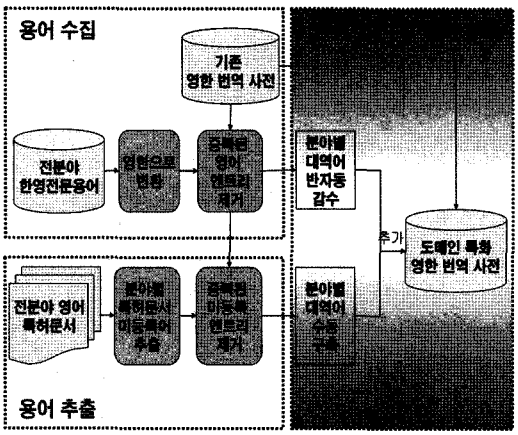


그림 1 용어 수집, 추출 및 구축

영한 특허 자동번역 시스템용 번역사전을 만들기 위해 대역 전문용어를 추가하는 방법은 두가지로 이루어질 수 있다. 하나는 기존의 역방향의 전문용어 사전이 존재할 때(예: 한영 전문용어 사전) 이를 활용하여 그림 1에서의 '분야별 대역어 반자동 감수' 방법과 같이 대역어를 반자동으로 구축하는 방법이다. 즉 대용량의 한국어 엔트리에 대한 영어 대역어가 부착되어 있는 대용량의 한영 전문용어를 영어 어휘를 표제어로 하여 영한으로 자동으로 변환한 후, 기보유 영한 번역 사전의 영어 표제어와 중복되는 영어 엔트리를 제거하고 남은 엔트리들에 대해 분야별로 대역어 반자동 감수를 통해 영한 전문용어화 시키는 작업이다.

다른 하나는 대량의 특허 문서가 존재할 때(예: 영어 특허 문서) 그림 1에서의 '분야별 대역어 수동 구축' 과 같이 미등록어를 자동으로 추출하여 한영 전문용어를 영한 전문용어로 변환한 엔트리와 중복되는 엔트리들을 제거한 후 고빈도순으로 대역어를 수동으로 구축 방법이 있다.

이상의 두가지 방법에 의해 총 1,840,235개의 신규 전문용어 엔트리가 구축되었는데 각 방법에 의해 구축된 엔트리 수는 다음과 같았다:

표 2 대역어 반자동 감수에 의해 구축된 용어수

구분	개수
전분야 한영 전문용어	3,052,655
영한으로 변환된 영어어휘	2,292,527
기존 영한 사전 엔트리수	836,000
기존 영한 사전 엔트리와 중복되지 않는 영어 어휘	1,726,571
1,726,571 중에서 구축된 용어 수	801,046(단일어:207,329, 빈도1이상의 복합어:593,717)

표 3 대역어 수동 구축에 의해 구축된 용어수

구분	개수
전분야 영어 특허문서	1,001,419
미등록어 추출	9,662,266(빈도1이상단일어(2,225,871), 빈도10이상 복합어(7,436,395))
9,662,266중에서 구축된 용어 수	1,039,189(빈도6이상단일어(492,295), 빈도91이상 복합어(546,894))[6]

3.2 특허 고유의 구문/문장 패턴 구축

앞서 2.4절에서 언급하였듯이, 특허 문서에서는 특허 고유의 문장 스타일을 보이는 구문 및 문장 패턴이 존재한다. 이러한 특허 고유의 구문/문장 패턴은 특허 전문번역가의 오랜 번역 경험 지식을 빌리지 않고는 구축할 수 없기 때문에, 특허 전문번역가에게 특허 문서에서 3단어 이상의 연속된 문자열을 고빈도순으로 정렬한 후, 그것이 포함된 문장을 특허 전문번역가가 보면서 특허 고유의 구문/문장 패턴을 수동으로 구축하는 방법을 취하였다. 수동으로 구축된 특허 고유의 구문/문장 패턴은 다음과 같았다:

5) 본 논문에서는 영한 특허문서 자동번역 시스템을 대상으로 하므로, 기보유 용어 수집의 대상은 영한의 역방향인 한국어 표제어에 대한 영어 대역어로 이루어진 한영 대역 전문용어로 그 대상을 한정하고자 한다.

6) 단일어가 빈도 6이상, 복합어가 빈도 91 이상 선택되어 수동 구축된 이유는 국가로부터 지원받은 구축 비용과 시간에 맞추어야 하는 문제가 있어 빈도 6이상, 빈도 91이상으로 한정을 시킨 결과이다.

표 4 특허 고유의 구문/문장 패턴 구축수

구분	개수
구문패턴	27,532
문장패턴	986
계	28,518

다음은 구축된 구문 및 문장 패턴의 예들이다.

- 구문 패턴의 예

  KEY: @NP_NP_as_claimed_in_claim_NUM

  CONTENT: { NP1! as claimed in claim NUM1 } -> { 제 NUM1 항의 NP1! }

- 문장 패턴의 예

  KEY: @S_NP_COMMA_wherein_S

  CONTENT: { NP1 COMMA wherein S1! } -> { NP1:[에] 있어서 COMMA S1! }

- 구문/문장 패턴에 의해 번역된 예

  원문: An air conditioner as claimed in claim 3, wherein the pump part includes;

  번역문: 제 3 항의 에어콘에 있어서, 펌프 부분은 다음을 포함한다 ;

3.3 형태소 태거의 특화

2.1 절에서 언급한 특허 문서의 특성을 근거로, 일반 도메인의 형태소 품사 태거를 특정 도메인에 적합한 형태소 품사 태거로 특화하는 3가지 방법을 다음과 같이 정의한다.

- 도메인 고유의 어휘표층형 특화: 특정 도메인에 발생하는 어휘 표층형에 대한 형태소 분석 관점에서의 특화
- 도메인 어휘통계정보 특화: 특정 도메인에 특이하게 발생하는 품사별 어휘 통계 또는 어휘별 품사통계 특화
- 도메인 문맥통계정보 특화: 특정 도메인에서 특이하게 발생하는 문맥통계(n-gram) 정보 특화

먼저, 특정 어휘 표층형에 대한 분석을 특화하기 위해서, 영어 토큰 분리와 형태소 분석 모듈을 부분적으로 수정하였다. 공백 단어 사이의 스트링들을 형태별로 분류하여서 분리할 심볼 및 대소문자 조합과 결합할 심볼 및 대소문자 조합을 구분하여 특허용 토큰 분리 모듈을 구현하였다. 그리고, 형태소 분석에서 사전에 등록되지 않은 "/"과 "-"으로 이루어진 복합 단어를 처리할 수 있도록 하였다.

특허 영한 자동번역 시스템의 품사 태거는 어휘화된 HMM(Hidden Markov Model)[7]을 확장하여 구현되었다. 그러므로, 어휘와 문맥통계 정보 특화 방법은 어휘 확률과 전이확률을 특허 도메인에 맞도록 조정하여야 한다. 특허 도메인용 어휘확률과 전이확률을 학습할 태깅된 특허 문서 코퍼스가 없으므로, 앞서 언급한 2001년부터 2005년까지의 약 100만 특허 문서를 일반 도메인용 태깅 시스템으로 태깅한 후, 일반 도메인에서 얻은 어휘확률과 0.3 이상 차이가 나는 어휘들을 수집하여 반자동으로 그 어휘확률을 조정하였다.

그리고, 또한 자동 태깅된 코퍼스에서 기존 어휘 및 품사 n-gram 통계치와 크게 차이가 나는 어휘 및 품사 n-gram을 전문가에 의해서 재조정하고, 새롭게 뽑힌 n-gram 정보를 기존 n-gram 통계치에 대비하여 추가하여 새로운 도메인 언어모델에서도 정확하게 품사 태깅이 이루어지도록 특화하였다.

3.4 구문 분석기의 특화

특허 도메인에 대해 구문분석기를 특화하는 주요 내용은 다음과 같다:

- 특허 고유의 구문/문장 패턴의 적용: 특허 고유의 구문/문장 패턴의 일반적인 형태는 어휘에 의해 구분되는 구문의 형태로 되어 있다. 이러한 패턴에 대해 패턴의 어휘를 먼저 인식하고, 인식된 패턴 가중치 순으로 패턴의 구문 노드에 대한 실제적인 파싱을 수행하여, 모든 구문 노드의 파싱이 성공하는 첫번째 패턴을 선택함으로써, 구문 분석 효율성 및 번역의 품질을 올렸다.
- 장문처리를 위한 병렬구조 인식 수행: 병렬구조 인식은 먼저 병렬구조의 시작점, 중간지점, 끝점이 될 수 있는 모든 가능한 지점을 인식한 후, 각 지점들간의 노드 유사도 테이블을 구한다. 그리고 모든 가능한 병렬 구조에 대해, 유사도 테이블을 이용해 최대의 병렬 가중치를 갖는 병렬 구조를 선택하여 구조분석이 성공하게 된다. 병렬구조가 인식되면, 이러한 병렬구조는 하나의 단위로 청킹이 됨에 따라, 문장이 단순화된다.
- 장문처리를 위한 문장 분할 수행: 병렬구조가 인식되고 난 후에도 구문분석기에서 한번에 처리하기에 너무 긴 문장인 경우에는 다시 문장 분할을 수행하는데, 문장분할은 분사구 나열이나 문장의 나열인 경우를 인식하는 것으로 시작점 패턴과 본동사 또는 분사의 유무 등의 조건을 체크하여 이루어진다.
- 전치사구, 분사구의 부착 편향성 반영: for 전치사구, 분사구에 대해 NP부착이 VP부착보다 우선하도록 편향성을 반영한다.

3.5 변환/생성기의 특화

일반 도메인에서 어휘들을 올바로 번역하기 위해 다양한 정보들(예를 들어, 공기정보, 격정보, 의미코드 등)을 사용하여 의미 모호성 오류 및 대역어 선택 모호성을 해소하였다.

일반 도메인 문서와 비교할 때 특허 문서에서의 대역어 선택 모호성의 심각성은 비교적 덜하다. 하지만, 특허 분야 간에 대역어가 달라질 수 있으며, 동일한 특허 분야 내에서는 대역어 선택 모호성 문제가 발생하기 때

문에 특허 문서 번역을 위한 변환기의 특화 작업을 수행하였으며 그 내용은 다음과 같다.

- 분야간 대역어 선택 해결을 위해 고빈도 어휘에 대한 특허 분야별 디폴트 대역어 등록: 동일한 영어 어휘가 서로 다른 특허 분야에서 서로 다른 한국어 대역어로 번역되는 경우, 대역어에 대해 특허 분야 문서별로 어휘의 고빈도를 추출하여 대역어에 대해 디폴트 대역어 정보를 자질값으로 할당하였다.
- 고빈도 명사/동사 어휘에 대한 대역어별 공기정보 수집: 동일한 특허 분야 내에서 문맥에 따라 서로 다르게 번역되는 어휘들을 수집하고, 이런 어휘들의 대역어 선택 모호성을 해결하기 위해 서로 다른 대역어로 번역되는 경우의 공기 정보들을 수집하였다. 이러한 공기정보는 대역어 선택 모호성을 해소하기 위한 실마리로서 사용된다.
- 동일 분야에서 공기정보를 활용한 대역어 선택 모호성 해소: 수집한 공기정보를 활용하여 동일 특허 분야 내에서 서로 다른 대역어로 번역되는 어휘들에 대한 대역어 선택 모호성 처리 모듈을 구현하였다.

생성 모듈의 경우, 기존의 일반 문서 번역을 위한 자동번역시스템과 비교해서 종결어미를 포함한 어미 등을 특허에 맞도록 수정하였으며, 특별한 특화 과정은 없었다.

4. 평가

본 절에서는 영한 특허 자동번역 시스템의 특허 분야별 번역률 평가 결과를 기술하고자 한다. 특허 분야별 평가 결과를 비교하기 위해 사용한 평가 코퍼스, 평가 방법, 평가 기준을 기술하면 다음과 같다:

- 평가 코퍼스: 2001년-2005년 사이에 출원된 100만여 건의 특허문서에서 주요 5개 산업분야[7](기계, 전기전자, 화학일반, 의료위생, 컴퓨터)에 대해 각 분야별로 임의로 1,000개의 문서를 선정하고, 선정된 문서들로부터 Field별 문장수와 가중치를 반영하여 각 분야별로 100 문장을 자동 추출하였다(분야별로 100문장을 구성하는 각 필드별 추출 문장수는 다음과 같다: Title(1문장), Abstract(2문장), Technical Field(1문장), Background of the Invention(5문장), Summary of the Invention(9문장), Brief Description of the Drawings(4문장), Description of the preferred Embodiments(54문장), Claims(24문장)).
- 평가 방법:
  - 3인의 특허 번역 전문가에게 정확성에 관한 스코어링 기준을 교육한 후 평가 기준에 따라 각자 평가 점수를 부여하게 하고 3인 평균으로 번역률을 계산함.
  - 번역률 산출방법은 다음과 같다:
    - 번역률(%)=개인별_번역률의_합(%)/평가자수
    - 개인별_번역률(%)=(개인별_총점/만점)×100
    - 만점 = 문장수×4점
- 평가 기준:

표 5 평가용 점수 부여 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 1-2단어의 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

위의 평가 기준에 의해서 특허 전문 번역가들이 점수를 부여한 실례를 보이면 다음과 같다:

[점수] 4점
[원문] A base station for a wireless network uses one or more MIMO channels having subchannels, to communicate with multiple user equipments, and allocates the sub channels to the user equipments.
[자동번역문] 무선 통신망을 위한 기지국은 다중사용자 장비와 통신하기 위해, 서브 채널을 가지는 하나 이상의 MIMO 채널을 이용하고, 사용자 장치에 서브 채널을 배정한다.

[점수] 3.5점
[원문] When the executable program is started, the loading system 1000 reads the started executable program 1810 from the memory 1800 into a work area of the loading system 1000 (step S61).
[자동번역문] 실행 가능 프로그램이 작동될 때, 로딩·시

7) 특허문서의 분야 분류는 국제적으로 표준이 되어 있는 IPC(International Patent Classification)을 따르는데 그 분류수가 59,759에 이르고 있다. IPC 분류가 너무 세분화되어 있어 산업적으로 직관적으로 분류할 수 있도록 국내 특허청에서는 12개의 산업분류법을 사용하고 있는데, 본 논문에서는 이 12개의 산업분류법을 따르도록 하였다. 12개의 산업분류법은 다음과 같다(전기전자, 화학일반, 기계, 의료위생, 컴퓨터, 채광금속, 농림수산, 섬유, 음료식품, 잡화, 토목건설, 사무용품). 이 12개의 산업분류중에 특허문서 출원수로 전기전자, 화학일반, 기계, 의료위생, 컴퓨터가 가장 많이 출원되어 있기 때문에 본 논문에서의 평가 분야로 5개 분야를 선택하였다.

스템 1000은 (S61 단계) 메모리 1800에서 로딩·시스템 1000의 워크 에어리어로 시동된 실행 가능 프로그램 1810을 리딩한다.
[설명] 문장 전체의 의미가 큰 왜곡없이 전달되므로 3.5점

[점수] 3점
[원문] Note that, in the present embodiment, a case in which the header information of the layer 4 protocol is used as specific information has been described.
[자동번역문] , 본 발명의 실시예에서, 층 4 프로토콜의 헤더 정보가 특정 정보로서 이용된 경우가 기술한다고 주목하라.
[설명] 'has been described가 '기술되었다는 것을'으로 번역되어야 옳으나, 문장의 전체적인 골격이 전달되므로 3점

[점수] 2.5점
[원문] A further variable criteria that can be used to formulate QoS decision factor thresholds to switch between the use of peer-to-peer techniques or infrastructure network communications is the volume of communication traffic in a particular geographic area serviced by one or more network base stations.
[자동번역문] QoS 결정 요인을 공식화하기 위해 이용될 수 있는 더 많은 가변성의 기준은 피어투피어 기술의 사용 사이에 전환하기 위해 스레스홀드하거나 기반 시설 망 통신은 하나 이상의 네트워크 기지국에 의해 서비스처리된 특별한 지리적인 영역에서 통신 트래픽의 볼륨이다.
[설명] 위 문장은 복문이고 밑줄친 것과 같은 오류가 있으나, 부분적으로 문장의 의미를 전달하므로 2.5점

[점수] 2점
[원문] Traditional media server solutions do not have a built-in admission control policy (for controlling the admission of new client requests to be serviced by the media server) that can prevent server overload and guarantee a desired quality of service.
[자동번역문] 전통적 배지 서버 솔루션이 내장된 인증 제어 방책을 가지지 않고 (제어하기 위해 새로운 클라이언트의 승인이 미디어 서버에 의해 서비스 처리되기 위해 요구한다) 서버 과부하를 방지하고 바람직한 서비스 품질을 보장할 수 있는.
[설명] 하나 이상의 구가 정확하게 번역은 되지만 문장 전체의 의미를 파악하기는 어려우므로 2점

[점수] 1점
[원문] As the above-mentioned solvent, there can be mentioned, hydrocarbon solvents such as benzene, toluene and the like;
[자동번역문] 상기 용매 언명할 수 있고, 탄화수소 용매.
[설명] 몇개의 단어 만이 번역되었고, 구단위 혹은 문장 단위 번역이 이루어지지 않음

위와 같은 평가 기준에 따라 100문장의 평가 코퍼스에 대해 영한 특허 자동번역 시스템을 평가한 결과는 다음과 같았다:

표 6 번역률 평가 결과 (평가일:2006.8.31)

구분	분야	문장당 평균단어수	번역률	
			전체	체감
영한특허 번역기	기계	32.38	80.54%	81.00%
	전기전자	30.11	81.58%	85.00%
	화학일반	27.36	79.92%	79.67%
	의료위생	28.77	80.79%	82.67%
	컴퓨터	27.19	82.29%	83.00%
평균		29.16	81.03%	82.27%

도표 6에서 전체번역률은 0~4점까지의 평가 점수의 총점에 대한 번역률을 말하며, 체감번역률은 평가 기준에서 3점 이상의 문장수에 대한 번역률을 말한다. 체감 번역률은 개발자 기준보다는 일반 사용자 입장에서 자동 번역 결과를 초벌번역용으로 활용할 수 있는 지의 수준을 수치화 한 것으로 번역 결과가 이해가 되느냐 안되느냐로 크게 구분하는 번역률이라 할 수 있다.

5. 결 론

본 논문에서는 일반 도메인을 대상으로 한 영한 자동번역기를 특허 도메인을 대상으로 한 영한 자동번역기로 특화하는 방법에 대해 살펴보았다. 특허 영한 자동번역기로의 특화 절차는 다음과 같은 절차로 이루어진다: 1) 대용량 특허 문서의 수집 및 언어학적 특성 분석, 2) 대용량 특허문서를 대상으로 한 전문용어 추출 및 대역어 구축, 3) 기존 번역사전 대역어의 특화, 4) 특허문서 고유의 번역 패턴 추출 및 구축, 5) 언어학적 특성 분석에 따른 번역 엔진 모듈의 특화 및 개선, 6) 특화된 번역 지식 및 번역 엔진 모듈에 따른 번역률 평가.

이러한 특화 방법에 의해 구현된 영한 특허 자동번역 시스템은 전분야 번역률 81.03%의 성능을 보이고 있다.

현재 본 논문에서 기술된 영한 특허 자동번역 시스템은 산업자원부의 특허지원센터에서 변리사 및 특허 심사관을 대상으로 전기전자분야 특허 문서에 대한 영한 특허 번역 서비스를 제공하고 있다.(http://www.ipac.or.kr) 또한 현재 전분야 특허문서에 대한 전문용어 및 번역 엔진 개선이 계속 진행 중에 있으며 2007년에는 전분야 특허문서에 대한 영한 자동번역 서비스를 제공할 예정이다.

본 논문에서는 다루어지지 않았지만 현재 추진되고 있으며 가까운 장래에 완성될 연구 주제로는 1) 영어 원문에 대한 10개의 Reference 정답셋을 만든 후 BLEU score에 의한 번역률 자동평가, 2) 특허 코퍼스를 기반으로 한 기존 번역 사전의 특허분야별 대역어 선택의 개선, 3) 철자 오류, 띄어쓰기 오류등과 같은 형태적 오류를 수정할 수 있는 저작 도구 개발, 4) 고품질의 번역률 제공을 위한 특허 문서 원본에 대한 통제 언어(Controlled language) 도구들이 연구 중에 있다.

참 고 문 헌

[1] 이민행, 지광신, 정소우 (1998), "기계번역 시스템 측정 장치 연구", 언어와 정보, Volume2, Number2.
[2] 시정곤, 김원경, 고창수 (2000), "영-한 기계번역 성능 평가 방안 연구", 언어와 정보, Volume4, Number2.
[3] 최승권 (2000) "영한자동번역에서의 두단계 영어전산문법", 언어와 정보, Volume1, Number1. 97-109쪽.
[4] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa and Makoto Iwayama (2003), "Patent Claim Processing for Readability - Structure Analysis and Term Explanation," ACL-2003 Workshop on Patent Corpus Processing.
[5] Remi Zajac (2003), "MT Customization," MT Summit IX Workshop.
[6] Munpyo Hong, Young-Gil Kim, Chang-Hyun Kim, Seong-Il Yang, Young-Ae Seo, Cheol Ryu, and Sang-Kyu Park (2005), "Customizing a Korean-English MT System for Patent Translation," MT Summit X. 181-187.
[7] Ferran Pla and Antonio Molina (2005), "Improving Part-of-speech Tagging Using Lexicalized HMMs," Natural Language Engineering 10(2) 167-189.

부록1. 컴퓨터 분야 영어특허문서 요약문의 자동번역 결과

(영어가 특허 원문이며, 한국어가 자동번역된 결과이다)

TITLE
제목
Systems and methods for presenting an interactive user interface
상호 작용하는 사용자 인터페이스를 제공하기 위한 시스템과 방법
ABSTRACT
요약
A system interactively controlled by a TV viewer remote control transmitter displays portions of a scroll program guide on the viewers display screen.
TV 뷰어 원격 송신기 제어에 의해 상호 반응적으로 제어된 시스템은 뷰어 디스플레이 스크린 위의 스크롤 프로그램 가이드의 부분을 디스플레이한다. A tuner receives TV radio frequency or optical transmission signals in a plurality of cable channels and passes a viewer usable signal to a signal combiner.
튜너는 다수의 케이블 채널에서 TV 무선 주파수 또는 광학 전송 신호를 수신하고, 신호 합성부로 뷰어 사용가능한 신호를 보낸다.
A computer receives control signals from the TV viewer remote control transmitter.
컴퓨터는 TV 뷰어 원격 송신기 제어로부터 신호 제어를 수신한다.
It controls the tuner to pass the viewer usable signal in response to one of the control signals.
그것은 제어 신호의 하나에 응하여 뷰어 사용가능한 신호를 통과하기 위해 튜너를 제어한다.
It receives and stores a scroll input picture image signal containing local program guide data and generates a scroll output picture image signal consisting of at least a portion of the scroll input picture image signal.
그것은 지역 프로그램 가이드 데이타를 포함하는 스크롤 입력 픽처 이미지 신호를 수신하고 보관하고, 스크롤 입력 픽처 이미지 신호의 최소한 부분으로 구성되는 스크롤 출력 픽처 이미지 신호를 발생시킨다.
The signal combiner combines the viewer usable signal from the tuner with the output picture image signal from the computer to provide a display signal for input to the viewers display screen.
신호 합성부는 입력을 위한 디스플레이 신호를 제공하기 위한 컴퓨터에서부터 뷰어 디스플레이 스크린까지 출력 픽처 이미지 신호와 튜너로부터의 뷰어 사용가능한 신호를 맞춘다.
The computer is responsive to variable control signals from the remote to advance, back up, and freeze the scroll output picture image signal.
컴퓨터가 원격 내지 전진의 가변성의 제어 신호에 응답하고, 뒤로 위로, 그리고 스크롤 출력 픽처 이미지 신호

를 결빙시킨다.

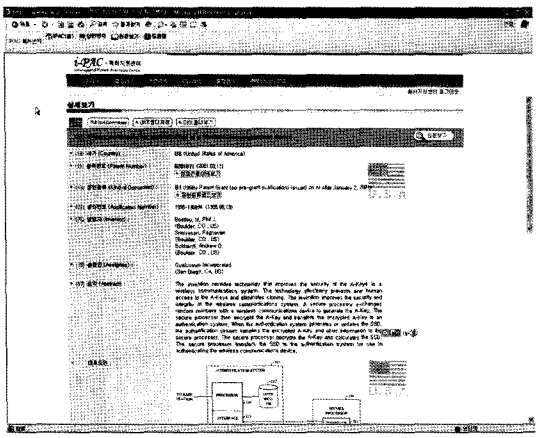
It is also responsive to directional control signals from the remote to reposition a highlight background to corresponding program data slots on the scroll grid and to display further program information corresponding to the program of the data slot shown in highlight.

스크롤 그리드 위의 상응하는 프로그램 데이타 슬롯에 하이라이트 배경을 옮기고 하이라이트에서 도시된 데이타 슬롯의 프로그램에 해당한 더 많은 프로그램 정보를 디스플레이하는 것이 원격으로부터 또한 방향적 제어 신호에 응답한다.

In addition, it is responsive to further directional control signals to redraw the grid to display earlier or later time segment program data than is normally displayed on the viewers screen.

게다가, 그것은 뷰어 스크린에 정상적으로 디스플레이되는 것보다 이르거나 더 늦은 시간 분절 프로그램 데이타를 디스플레이하기 위해 그리드를 재작성하기 위해 더 많은 방향적 신호 제어에 응답한다.

부록2. 산업자원부 산하 특허지원센터의 전기전자 영한 특허문서 자동번역 서비스 화면



최 승 권

1987년 한국외대 독일어과 졸업(학사) 1992년 독일 빌레펠트대 전산언어학과 졸업(석사). 1995년 독일 자아란트대 자동번역과 졸업(박사). 1995년~현재 한국전자통신연구원 언어처리연구팀 책임연구원. 관심분야는 자동번역, 자연어처리

권 오 욱

1992년 경북대학교 컴퓨터공학과(학사) 1995년 KAIST 전산학과(석사). 2001년 포항공대 컴퓨터공학과(박사). 2004년~현재 한국전자통신연구원 언어처리연구팀 선임연구원. 관심분야는 자동번역, 정보검색, 문서분류

이 기 영

1994년 한양대학교 컴퓨터 공학과(학사) 1997년 한양대학교 컴퓨터 공학과(석사) 2000년~현재 한국전자통신연구원 언어처리연구팀 선임연구원. 관심분야는 자동번역, 정보검색, 영상처리

노 윤 형

1997년 KAIST 전산학과(학사). 2000년 KAIST 전산학과(석사). 2000년~현재 한국전자통신연구원 언어처리연구팀 연구원. 관심분야는 자동번역, 자연언어처리

박 상 규

1982년 서울대학교 컴퓨터공학과(학사) 1984년 KAIST 전산학과(석사). 1997년 KAIST 전산학과(박사). 1987년~현재 한국전자통신연구원 언어처리연구팀 책임연구원. 관심분야는 자동번역, 정보검색