제한된 도메인에 특화된 기계번역 기술 개발
- 특허 전문 영한 번역기를 중심으로 -

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

Development of Machine Translation Technology Customized at Restricted Domain - Focusing on English-Korean Patent Translator -

Sung-Kwon Choi, Eunjin Park, Young-Kil Kim
NLP Research Team, ETRI

요 약

본 논문은 2005 년부터 2006 년도까지 정보통신부의 지원 하에 한국전자통신연구원 언어처리연구팀에서 성공적으로 개발하여 현재 산업자원부 특허지원센터에서 대용량의 영어 특허문서를 대상으로 한국어 자동번역 서비스를 제공하고 있는 특허 전문 영한 번역기에 대해 기술한다. 특히 본 논문에서는 일반 도메인을 대상으로 한 기존의 영한 번역기를 제한된 도메인을 대상으로 한 영한 번역기로 개량하고자 할 때, 개량하는 방법으로써 제한된 도메인에 대한 특화 절차에 대해서 기술한다. 이와 같이 특화 절차에 따라 구축된 특허 전문 영한 번역기 번역률을 특허 분야 중에 주요 5 개 분야(기계, 전기전자, 화학일반, 의료위생, 컴퓨터)에 대해 특허전문번역가가 평가한 결과, 평균 82.43%가 나왔다. 또한 전기전자 분야 특허문서를 대상으로 특허 전문 영한 번역기와 일반 도메인을 대상으로 한 영한 번역기와의 번역률을 평가한 결과, 특허 전문 영한 번역기는 82.20%, 일반 도메인 대상 영한 번역기는 54.25%의 번역률을 내어, 특허에 특화된 특허 전문 영한 번역기가 특화되지 않은 일반 도메인의 영한 번역기에 비해 27.95%나 더 높은 결과를 알 수 있었다.

1. 서론

각국의 특허청에서는 특허기술의 신속한 권리와 보호를 위해 특허 출원 및 등록 심사 처리 기간을 단축하려는 노력을 하고 있다. 이러한 노력의 일환으로 특허행정 정보화 시스템의 고도화가 이루어지고 있다.

특허행정 정보화 시스템의 고도화와 관련해 특허청에서는 특허 자동번역 시스템을 이용한 특허 자동번역 서비스를 실시함으로써 심사관 및 특허 출원자들이 선행기술조사를 더욱 편리하고 빠르게 수행할 수 있도록 지원하고 있다.

본 논문에서는 2005 년도부터 2006 년도까지 정보통신부의 지원 하에 한국전자통신연구원 언어처리연구팀에서 성공적으로 개발하여 현재 산업자원부 특허지원센터(http://www.ipac.or.kr)에서 대용량의 영어 특허문서를 대상으로 한국어 자동번역 서비스를 제공하고 있으며, 가까운 시일 내에 특허청에서도 영한 특허문서 자동번역 서비스를 개시하게 될 특허 전문 영한 번역기를 기술하는 것을 목표로 한다.

2. 제한된 도메인과 번역 품질

80 년대 초반부터 외국에서는 자동 번역기를 산업현장에서 직접 활용할 수 있는 방안을 모색하기 시작하였다. 그 결과 자동 번역기의 대상을 일반 도메인에서 항공기 매뉴얼, 섬유, 컴퓨터, 주식 시장 보고서, 핸드폰 메뉴얼과 같은 제한된 도메인으로 특화하여 적용하는 시도를 하였다. 그 결과 일반 도메인보다도 상당히 유용한 번역 결과를 생성해 낼 수 있었다.[1] 그 이유는 제한된 도메인의 문서가 일반 도메인에서의 문서보다 자동번역을 하는데 있어서 다음과 같은 장점을 가지기 때문에 가능했다: 1) 제한된 도메인에는 동형이의어(homography)가 나타나면 하나의 특정 의미로 사용된다 2) 제한된 도메인에는 대역어 선택 모호성이 일반 도메인에서보다 덜하다 3) 제한된 도메인에는 구조적 편향성 때문에 구조적 모호성이 상당 부분 해결된다.

결과적으로 자동번역 시스템의 태생적인 문제인 모호성 문제들이 제한된 도메인에서는 어느 정도 해소될 수 있기 때문에, 일반 도메인에서보다는 제한된 도메인에서 자동번역 시스템의 번역률이 상승될 수 있다. 그러나 문제점 또한 존재하는데, 제한된 도메인의 문서들이 전문 분야여서 대량의 전문용어를 인식 및 구축해야 하는 문제와, 전문 문서에 등장하는 원문의 형태적, 통사적, 의미적 특성과 고유한 패턴을 시스템에 반영하여야 하는 문제는 여전히 존재한다.

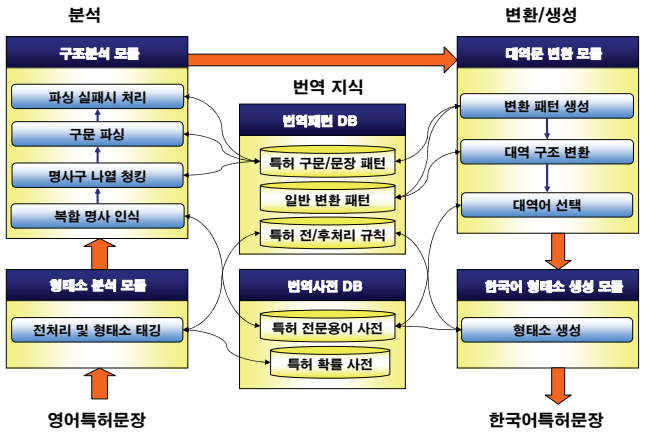
3. 제한된 도메인으로의 특화 절차

번역기를 대상으로 일반 도메인으로부터 제한된 도메인으로 특화하는 방법으로, 다국어 자동번역 시스템인 SYSTRAN 을 대상으로 하는 특화 방법[2], 특허문서를 대상으로 한영 자동번역기를 특화하는 방법[3], 특허문서를 대상으로 영한 자동번역기를 특화하는 초기 방법[4][5]이 소개된 바 있다.

영어 특허 문서를 대상으로 일반 영한 번역기로부터 특허 전문 영한 번역기로 특화하는 절차는 다음과 같다: 1) 대용량 영어 특허 문서에 대한 언어적 특성 분석, 2) 대용량 영어 특허 문서를 대상으로 한 전문용어의 자동 추출 및 반자동 대역어 구축, 3) 기존 전문용어 사전의 대역어 튜닝, 4) 영어 특허 고유의 특허 번역 패턴 자동 추출 및 반자동 대역 패턴 구축, 5) 언어적 특성 분석에 따른 번역 엔진 모듈의 특화 및 개선, 6) 특화된 번역 지식 및 번역 엔진 모듈에 따른 임의의 특허 문서 대상 번역률 평가.

4. 특허 전문 영한 번역기의 시스템 구성도

특허 전문 영한 번역기는 크게 번역 엔진과 번역 지식으로 나뉜다. 번역 엔진은 4 개의 모듈로 구성되며, 번역 지식은 2 개의 DB 로 구성된다. 4 개의 모듈로 구성되는 번역 엔진은 특허에 특화된 형태소 분석 모듈, 특허에 특화된 구조 분석 모듈, 특허에 특화된 대역문 변환 모듈, 특허에 특화된 한국어 형태소 생성 모듈이며, 2 개의 DB 로 구성되는 번역 지식은 특허 전문 번역 패턴 DB 와 특허 전문 번역 사전 DB 이다. 이와 같은 특허 전문 영한 번역기의 시스템 구성도를 그림으로 보이면 다음과 같다:



(그림 1) 특허 전문 영한 번역기의 시스템 구성도

5. 번역률 평가

본 장에서는 특허 전문 영한 번역기의 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: 영어 특허 문서 100 만 여건 중에서 주요 5 개 산업분야(기계, 전기전자, 화학일반, 의료위생, 컴퓨터)에 대해 각 산업 분야별로 임의로 1,000 개의 문서를 선정하고, 각 산업 분야별로 100 문장을 개별적으로 자동 추출하였다.
- 평가 방법:
  - 7 인의 특허 전문 번역가에게 평가 점수 부여 기준을 교육한 후 평가 기준에 따라 각자 평가 점수를 부여하게 하고 각 문장별로 최고 최저점을 제외한 5 개 점수의 평균 합으로 번역률을 계산함.
  - 번역률 산출방법은 다음과 같다:
    - ◆ 변역률(%)=최고_최저점을_제외한_5 인의_번역률의_합(%)/평가자수_5 인
    - ◆ 개인별_번역률(%)=(개인별_총점/만점) x 100
    - ◆ 만점 = 문장수 x 4 점
- 평가 기준:

<표 1> 평가용 점수 부여 기준

점수	평가 기준
4.0	원어문의 의미가 그대로 전달됨
3.5	복문에서, 문장의 동사구가 정확히 전달되어 문장의 전체적인 의미의 골격이 전달되지만 동사를 제외한 1-2단어의 대역어가 잘못됨
3.0	문장의 동사구가 정확히 전달되어 문장의 전체적인 의미의 골격이 전달됨
2.5	하나의 동사절이라도 정확히 번역되어 부분적으로 문장의 의미를 전달함
2.0	하나 이상의 구가 정확히 번역되지만 전체적인 문장의 의미를 파악하기 어려움
1.0	문장 중에 하나의 단어 또는 구라도 정확히 번역됨
0.0	번역문 출력이 안 됨

위와 같은 평가 기준에 따라 각 분야별로 개별적으로 선정된 100 문장의 평가 코퍼스에 대해 5 개 분야에 대해 특허 전문 영한 번역기를 평가한 결과는 다음과 같다:

<표 2> 5 개 산업분야 특허문서의 번역률 평가 결과
(평가일:2006.12.13)

분야	문장당 평균단어수	전체번역률	체감번역률
기계	30.34	83.50%	85.00%
전기전자	28.19	82.20%	88.00%
화학일반	29.67	82.20%	91.00%
의료위생	26.75	81.63%	86.00%
컴퓨터	25.49	82.63%	88.00%
평균	28.09	82.43%	87.60%

<표 2>에서 전체번역률은 0~4 점까지의 평가 점수의 총점에 대한 번역률을 말하며, 체감번역률은 평가 점수에서 3 점 이상의 문장에 대한 번역률을 말한다. 체감번역률은 일반 사용자 입장에서 자동 번역 결과를 초벌번역용으로 활용할 수 있는 지의 수준을 수치

화 한 것으로, 번역 결과가 이해가 되느냐 안 되느냐의 O, X 로 크게 구분하는 번역률이라 할 수 있다.

<표 2>에서 나타난 바와 같이, 영어 특허문서의 주요 5 개 분야의 평균 번역률은 82.43%이며, 체감번역률은 87.60%로써, 특허 전문번역가가 특허 전문 영한 번역기의 번역결과를 초벌번역용으로 충분히 활용할 수 있는 수준이라는 것을 알 수 있다.

또한 다음의 <표 3>은 상기의 <표 2>에서 전기전자분야의 원문에 대해 특허에 특화된 영한 번역기와 특허에 특화되지 않은 영한 번역기의 번역 결과에 대해 비교 평가한 결과이다.

<표 3> 전기전자분야의 특화 전/후 번역률 비교 평가 (평가일:2006.12.13)

분야	문장당 평균단어수	특화 전 번역률	특화 후 번역률
전기전자	28.19	54.25%	82.20%

<표 3>에서 알 수 있는 것은 전기전자분야의 특허문서에 대해 특화전의 영한 번역기와 특화후의 영한 번역기의 번역률 차이는 27.95%가 난다는 것이며, 이것은 본 논문에서 제시된 특화절차가 특허와 같은 제한된 도메인에 대해서 번역품질을 향상시키는데 중요한 역할을 하였다는 것을 의미한다.

6. 결론

본 논문에서는 일반 도메인을 대상으로 한 영한 번역기를 특허 도메인을 대상으로 한 영한 번역기로 특화하는 절차 및 그에 따른 번역률 향상 결과를 살펴보았다. 특허 전문 영한 번역기로의 특화 절차는 다음과 같은 절차로 이루어진다: 1) 대용량 영어 특허 문서에 대한 언어적 특성 분석, 2) 대용량 영어 특허 문서를 대상으로 한 전문용어의 자동 추출 및 반자동 대역어 구축, 3) 기존 전문용어 사전의 대역어 튜닝, 4) 영어 특허 고유의 특허 번역 패턴 자동 추출 및 반자동 대역 패턴 구축, 5) 언어적 특성 분석에 따른 번역 엔진 모듈의 특화 및 개선, 6) 특화된 번역 지식 및 번역 엔진 모듈에 따른 임의의 특허 문서 대상 번역률 평가.

이러한 특화 절차에 의해 구현된 특허 전문 영한 번역기는 특허의 주요 5 개 분야에서 평균 번역률 82.43%의 성능을 보였으며, 특화 전 영한 번역기와 비교하여 전기 전자 분야에서 27.95%의 성능 향상이 있었다.

현재 본 논문에서 기술된 특허 전문 영한 번역기는 산업자원부의 특허지원센터에서 변리사 및 특허 심사관을 대상으로 영어 특허 문서에 대한 영한 특허 번역 서비스를 제공하고 있다.(http://www.ipac.or.kr).

참고문헌

[1] John Hutchins (1992). An Introduction to Machine Translation. Academic Press.

[2] Remi Zajac (2003) “MT Customization”. MT Summit IX Workshop.

[3] 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

[4] 최승권, 권오욱, 이기영, 노윤형, 박상규 (2006) “웹 영한 번역기로부터 특허 영한 번역기로의 특화 방법”. 제 18 회 한글 및 한국어 정보처리 학술대회, 포항공대. 57-64.

[5] 최승권, 권오욱, 이기영, 노윤형, 박상규 (2007) “도메인 특화 방법에 의한 영한 특허 자동 번역 시스템의 구축”. 정보과학회 논문지, 제 34 권 제 2 호, 95-103.