번역학연구
2014년 겨울 제15권5호

언어학적 지식에 기반한 한중 뉴스 표제의 기계번역

황 은 하
(배재대)

1. 서론

본 연구는 한중 뉴스 표제의 대조 분석을 토대로 기계번역에 적용 가능한 언어학적 지식을 규칙으로 정리하고, 이를 기계번역에 적용하여 뉴스 표제의 기계번역 정확률을 높이는 데 목적이 있다.

황은하(2009)에 따르면, 기계번역 기술은 의존하는 언어 지식의 유형과 그 활용 방법에 따라 크게 규칙 기반 기법(R.M.K. Sinha, 2002)과 말뭉치 기반 기법(Kenji Ono, 2003)으로 나뉜다. 후자의 경우 다시 예문 기반 기법(Diganta, S., Sivaji, B., 2005, John Hutchins, 2005), 패턴 기반 기법(황은하, 홍문표, 최승권, 2002) 등의 지식 기반 기법과 통계 기반 기법으로 나누어 볼 수 있다. 최근 들어 IT 기술과 빅데이터라는 개념의 도입 및 처리 기술의 빠른 발전에 힘입어, 기계번역에서는 통계 기반 번역 기법을 사용하는 것이 대세다. 구글 기계번역 팀의 책임자 베누고팔(Venugopal)은 “기계번역은 일종의 빅데이터 처리 시스템이다. 우리는 많은 데이터를 처리해 스스로 언어를 배우는 시스템을 만든다. 각

언어 학습은 시스템이 자동으로 한다. 그래서 우리 팀 대부분은 데이터 처리 기술 전문가이며, 언어학자는 1명도 없다. 언어 이해보다 데이터를 잘 이해하는 사람이 필요하다."[1]고 기계번역에 대한 관점을 밝힌 바 있다. 이처럼 언어학적 지식보다는 대규모 데이터에 기반한 통계 기반 기법이 기계번역 기술 개발 현장에서 상당히 우세하고 있는 것이 사실이다.

기계번역이 다루는 것은 언어와 언어 간의 번역 문제인데, 언어학적 지식은 정말 필요하지 않은 것일까? 본 연구는 이와 같은 문제 제기에서 출발하여, 현실적으로 기계번역의 필요성이 고조되고 있는 뉴스 표제를 대상으로 언어학적 연구를 통한 기계번역의 정확률 향상을 시도하고, 그 정도를 측정하여 언어학적 지식의 기계번역에 대한 기여도를 객관적으로 입증하고자 한다.

이를 위해 본 연구는 2장에서는 한중 기계번역기의 연구 현황을 살펴보고, 본 연구의 실험에 이용할 KC-Tran[2]의 시스템의 특징에 대해 살펴본다. 이어서 3장에서는 KC-Tran 시스템 구성에 적합한 형식으로, 뉴스 표제의 한중 기계번역에 적용할 수 있는 언어 지식 및 규칙을 정리한다. 4장에서는 3장에서 정리한 지식을 기계번역기에 적용하여 선언적 평가(Declarative evaluation)[3] 방법으로 기계번역 정확률의 향상 정도와 더불어 언어학적 지식의 기여도를 평가한다. 마지막으로 5장에서는 본 연구에 대한 요약과 더불어 본 연구의 의의 및 남은 과제에 대해 기술한다.

---

1) 2012년 11월 3일자 Chosun Biz에 실린 "[7 Questions] 구글 기계번역 책임자 베누고팔"이라는 표제의 인터뷰 기사에서 발췌했으며, URL 주소는 다음과 같다. http://m.biz.chosun.com/svc/article.html?contid=2012110201300

2) 본 연구의 평가 부분을 허락해 주시고, 4장의 번역 평가세트 100문장과 패러프레이징된 문장 100문장, 총 200문장에 대해 번역문을 생성해 준 한국전자통신연구원의 언어처리연구팀에 깊은 감사를 드린다.

3) 선언적 평가는 입력과 출력 모두에 명료성(Intelligibility), 정확성(Accuracy), 문체(Style) 등을 기준으로 삼는 기계번역 평가 방법의 하나이다. 4.1에서 보다 자세히 기술하도록 한다.

2. 한중 기계번역기 개발 현황

한국어와 중국어 간의 기계번역은 출발어와 목표어에 따라 중한 기계번역과 한중 기계번역으로 나뉜다. 중한 자동 번역은 기업체 및 학교를 중심으로, 중국 대학교 및 연구소와의 협업 하에 중국어 분석 기술을 확보하여 개발이 활발한 편이다. 한편, 한중 자동 번역 기술 개발은 한국어 분석 기술 및 중국어로의 변환 기술의 어려움으로 인해 다른 어종에 비해 뒤늦은 출발을 하였으나 최근 들어 여러 가지 웹 기반 한중 번역기가 개발되어 상용화되는 등 많은 성과를 거두고 있다.

한국어를 출발어로 하고 중국어를 목표어로 하는 한중 기계번역 시스템은 지금까지 알려진 바로, 마이크로소프트(Microsoft), 바빌론(Babylon)[4], 구글(google)[5], 유다오(有道)[6]와 한국전자통신연구원(ETRI)의 KC-Tran 등 다섯 가지가 있다. 이밖에 북한에도 성능이 뛰어난 조중(朝中) 기계번역 프로그램이 개발되었다고 전해지고 있으나, 공개된 문헌에서는 관련 내용을 찾을 수 없었다[7].

마이크로소프트, 구글, 바빌론, 유다오는 모두 웹 기반 기계번역 시스템으로, 일반에 공개되어 번역 서비스를 제공하고 있다. 시험적으로, 뉴스 표제 “한국, 100년후 아열대로”를 입력했을 때 상술한 시스템마다 각각 서로 다른 번역문을 출력해 보이고 있다. 다음은 기계번역기별로 번역 결과인 출력문과, 그에 대한 한국어 번역문을 붙인 것이다.

---

4) 마이크로소프트의 한중 기계번역기와 바빌론 한중 기계번역기는 모두 아임트랜슬레이터닷넷(http://wwww.imtranslator.net)에 접속하여 사용할 수 있다. URL 주소는 다음과 같다. http://imtranslator.net/translation/korean/to-chinese-simplified/translation

5) 구글 번역 시스템은 구글 홈페이지의 번역 채널(http://translate.google.com)에서 서비스되고 있다.

6) 유다오(有道) 는 중국 최대 포털사이트 왕이(網易) 에서 제공하는 웹 기반 기계번역 시스템이다. 통계기반 기법을 사용하고 있으며, 한중 기계번역 외에도 중국어와 일본어, 영어, 한국어, 불어 등의 쌍방향 기계번역을 제공한다. URL 주소는 http://fanyi.youdao.com 이다.

7) 이봉원(2002)에 따르면, 북한은 일찍이 1960년대부터 언어학자들이 수학, 전자공학 등 분야의 전문가들과 공조하여 기계번역 연구에 착수하였으며, 조일(WINKTRAN, 담징 1.0), 일조(해돋이, 대동문), 일영(무지개 2.2), 로조(대동문) 등의 언어쌍에 대한 기계번역 프로그램이 개발되었다고 한다.

〈표 1. 시스템별 한중 기계번역 출력문 비교〉

	기계번역 결과	출력문의 번역
마이크로소프트	100 年后到 亞熱帶韓國	100년후 아열대한국
바빌론	韓國, 100年之后的亞熱帶气候	한국, 100년 후의 아열대 기후
구글	韓國，100多年的亞熱帶	한국, 100여년의 아열대
유다오	100年后, 韓國以亞熱帶	100년후, 한국은 아열대로써

이상에서처럼, 동일한 입력문에 대해 네 가지 시스템 모두 각기 다른 번역문을 생성한 것을 알 수 있는데, 이는 시스템마다 기계번역 기법이 다르고 엔진을 구성하고 있는 사전과 언어 데이터베이스도 다르기 때문인 것으로 분석된다.

마이크로소프트의 번역은 구문분석 실패로 두 단락으로 나뉜 부정확한 출력문을 생성했고, 바빌론은 반점(,)을 나열 관계에 쓰이는 일반 텍스트의 반점과 같이 인식하여 번역문에서 모점(、)을 사용해 '한국'과 '100년 후의 아열대 기후'를 동등한 자격으로 만들어 줬으며, 표제 안의 문법 관계 구현에 실패했다. 구글의 번역문은 한국어의 반점을 그대로 반점으로 변환하여 생성했다는 점 외에는 바빌론의 번역문과 별반 다르지 않다. 유다오는 번역문에 '100년후'를 부사어로 정확히 분석하여 이를 표제의 맨 처음에 생성한 것으로 보이며, 다만 서술어가 생략된 부사어 '아열대로'에 쓰인 '-로'의 대역어 선택에 실패한 것을 알 수 있다. 입력문의 '-로'는 변화의 결과를 나타내는 데 반해, 출력문에서는 어떤 일의 방법이나 방식을 나타내는 '-로'의 의미인 '以'라는 잘못된 대역어를 출력한 것이다.

구글의 기계번역 시스템은 앞서 베누고팔의 인터뷰에서도 밝혀진 것처럼, "번역문을 생성할 때 사용자에게 가장 적합한 번역을 제공하기 위해 수억 개의 문서에서 패턴을 조사하며, 사람이 이미 번역한 문서에서 패턴을 조사함으로써 적절한 번역이 무엇인지에 대해 지능적으로 추측"[8]한다. 즉 대규모 말뭉치에서 입력문과 비슷한 패턴을 찾아내는 통계 기반 기계번역 기법을 사용하고 있다. 유다오 역시 주로 통계 기반 기계번역 기술을 사용하고 있는 것으로 알려진 데

8) 구글 번역 홈페이지의 "번역이 이루어지는 과정 알아보기" 페이지에서 인용하였으며, URL은 다음과 같다. http://translate.google.co.kr/about/intl/ko_ALL

반해, 마이크로소프트사와 바빌론의 한중 기계번역 시스템에 대해서는 알려진 바가 거의 없다.[9]

본 연구에서 실험에 적용할 한국전자통신연구원(ETRI)의 KC-Tran은 주로 동사 패턴(pattern)에 기반한 기계번역 방법을 채택하고 있으며 이는 동사구의 번역에서 보다 정확한 표현을 제공하고 동사의 의미 모호성을 해결하며 동사의 논항이 되는 명사의 중의성을 해소하는 데에도 결정적 영향을 미친다(황은하, 최승권, 홍문표, 2002).

KC-Tran 한중 번역기는 언어 지식 기반 기계번역 시스템으로, 크게 한국어 형태소 분석기, 동사 패턴 기반 구문분석기와 중국어 생성기 등 세 부분의 모듈로 구성되어 있다. 번역이 필요한 한국어 문장이 입력되면, 형태소 분석기에서 분석 사전과 전처리 규칙 등을 이용하여 형태소 단위로 끊어서 분석하며, 통계 정보를 이용하여 최적의 형태소 분석 결과를 선택한다. 형태소 분석 결과는 구문 단계의 분석으로 이어지며, 적절한 구문 단위로 묶여지게 된다. 다음 여러 개의 언어 지식 사전을 통해 목표 언어로 부분 변환하며 최종적으로 생성 모듈을 통해 중국어 출력문이 완성된다.[10]

3. 기계번역을 위한 한중 뉴스 표제의 지식

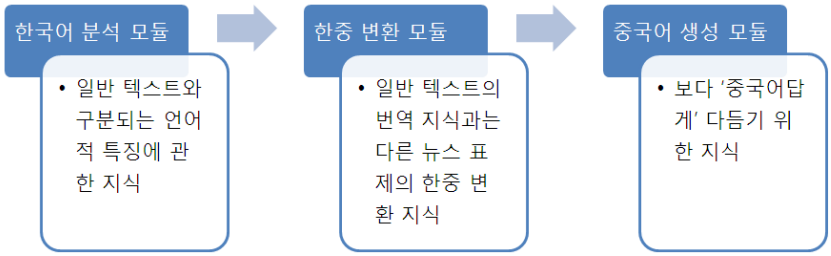
3장에서는 한중 뉴스 표제의 기계번역에 필요한 지식을 정리하고, 이를 기계번역 시스템에 어떻게 적용할지에 대해 논의한다.

앞서 2장에서 살펴본 것처럼 평가에 이용할 KC-Tran의 시스템 구성은 크게 한국어 분석 모듈, 한중 변환 모듈과 중국어 생성 모듈로 나뉜다. 이론적으로, 한국어 분석 모듈은 한국어 입력문, 즉 한국어 뉴스 표제의 일반 텍스트와 구분되는 언어 특징에 대한 지식을, 한중 변환 모듈은 일반 텍스트와 구분되는 한중 뉴스 표제의 대응 규칙을 필요로 하며, 중국어 생성 모듈은 중국어 뉴스 표제만

9) 마이크로소프트사와 바빌론처럼 연구 기관이 아닌 영리를 목적으로 하는 회사의 성격상 보유하고 있는 기술을 논문이나 보고서 등의 형식으로 공개하지 않는 것이 관례이다.

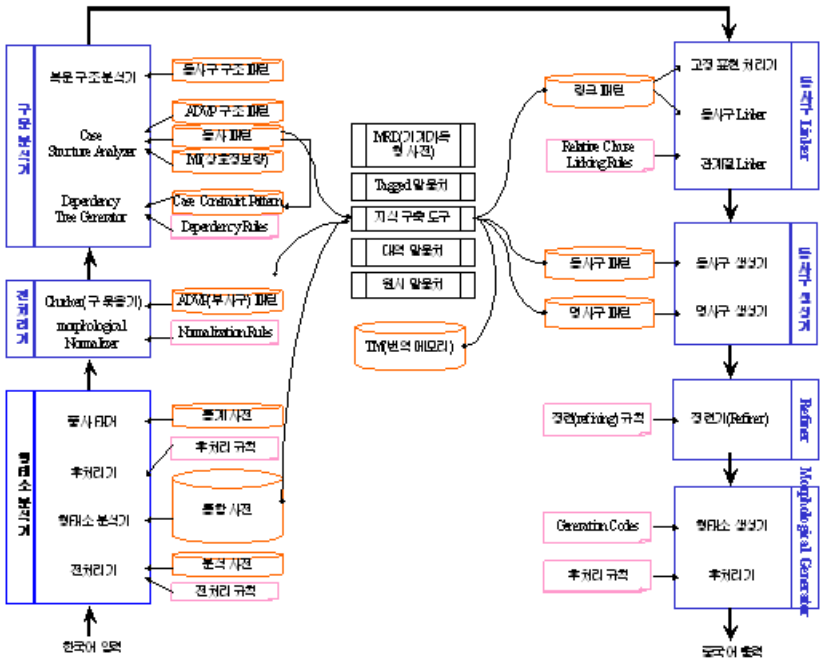
10) 다음은 황은하, 홍문표, 최승권(2002)에서 인용한 KC-Tran의 엔진 구성도이다.

의 언어 특징에 대한 지식을 필요로 한다. 각각의 모듈에 필요한 언어 지식을 그림으로 보이면 다음과 같다.



〈그림 1. 대조 분석 연구의 기계번역 응용〉

이처럼 언어 연구의 결과를 기계번역기에 적용 가능한 기계가독형 데이터로 정리하고 적용하는 일은 언어학과 언어공학이라는 상이한 두 학문이 기계번



역이라는 접점에서 만나는 이상적인 협업 모델임이 틀림없다. 그러나 현실적으로 여러 가지 어려움이 있다. 우선, 언어 연구의 결과를 기계번역 시스템의 엔진 특성을 고려하여 모두 기계가독형으로 정리하는 일은 결코 수월한 일이 아니다. 설령 기계가독형으로 모두 전환했다고 하더라도 신구(新舊) 언어 규칙과 지식이 서로 충돌하여 원래보다 못한 번역문을 출력하는 경우도 있다. 이뿐만 아니라, 위의 응용 모델대로라면 기계번역 시스템의 분석, 변환과 생성의 세 모듈을 모두 수정하고 개선해야 하며, 많은 시간과 노력을 요하게 된다. 따라서 본 연구는 실험 및 평가를 위한 지식의 적용에 필요한 수고를 최소한으로 줄이기 위해 패러프레이징(paraphrasing) 기법을 도입해 입력문을 고쳐 쓰는(rewrite) 방법을 채택한다.

3.1 지식의 적용: 패러프레이징(paraphrasing)

기계번역에서 말하는 패러프레이징은 입력문의 단어와 구 및 문장 중 일부를 분석 및 번역하기에 수월한 형태로 바꿔줌으로써 분석 및 번역의 정확률을 개선시키는 데 활용된다. 예를 들면 긴 복문을 여러 개의 단문으로 고쳐써서 구문분석 정확도와 변환 정확률을 높일 때 사용된다. 또는 원문에 나타난 잘못된 표현을 바르게 고쳐 쓰는 방법을 통해, 오류 또는 미등재어의 번역 오류를 줄이기도 한다. 이외에도 입력문의 생략된 주어 복원, 격 복원 등을 통하여 입력문에서는 드러나지 않은 격 관계를 번역문에서 표현하기도 한다.

패러프레이징 기법은 필요로 하는 지식의 유형에 따라 통계 기반과 규칙 기반의 두 가지 방법으로 나뉜다. 전자는 기계학습을 위한 대규모의 패러프레이징 전후의 대응 문장쌍 데이터를 필요로 한다. 본 연구에서는 후자, 즉 한국어 뉴스 표제의 문법적 및 어휘적 특성에 대한 분석, 변환, 생성 규칙에 기반하여, 뉴스 표제를 일반 텍스트의 형식으로 바꿔주기 위한 패러프레이징 규칙을 제시하고 적용하여 뉴스 표제의 기계번역 성능을 향상시키고자 한다.

3.2 한중 뉴스 표제 기계번역 응용 지식

여기서는 한중 뉴스 표제의 대조 분석 결과를 토대로 한국어 입력문에 대한 패러프레이징 규칙과 더불어 시스템에 대한 큰 수정이 필요 없이 적용 가능

한 몇 가지 지식을 기계번역기 KC-Tran의 모듈별로 나누어 적용 가능한 20가지의 규칙으로 정리하였다.

3.2.1 한국어 분석 모듈

1) 미등재 고유명사의 분석

뉴스 표제에는 미등재 고유명사가 많이 나타나며 기계번역의 한국어 분석 단계에서 실패할 수밖에 없다. 이를 위해 미등재 고유명사의 출현 환경에 대해 다음과 같은 규칙을 적용할 수 있다.

규칙 1. 미등재어이면서 표제 문두에 오거나 반점 앞에 오는 경우, 해당 형태소를 고유명사로 처리한다.

(1) 하이닉스, 반도체 장비 일부 중(中)업체에 매각키로[11)]
(2) 미네르바, 대선 때 MB 퇴진 운동 단체 회원 가입해 활동

(1)의 '하이닉스'와 (2)의 '미네르바'는 각각 사전 미등재어로, 표제의 문두에 오는 동시에 반점 앞에 위치하며, 규칙 1에 근거하여 고유명사로 정확히 분석될 수 있다.

2) 절의 분석

일반 텍스트에서 절과 절은 연결어미로 이어지며 쉼표로 그 구분을 뚜렷이 하는 것이 일반적이지만, 뉴스 표제에서는 절의 구분에 있어 말줄임표(.....)를 주로 사용하는 것으로 나타났다. 한편, "한글맞춤법"에 따라 말줄임표는 여섯 개의 온점을 찍는 것이 정확하지만, 뉴스 표제에서는 온점 2개(..), 온점 또는

11) 본 연구에 사용된 한국어 뉴스 표제 예문은 모두 2009년 조선일보, 중앙일보, 동아일보의 한국어와 중국어 번역문을 수집하여 구축한 황은하(2012)의 한중 뉴스 표제 병렬말뭉치에서 추출하였다.

가운뎃점 3개(..., …)를 포함해 다양한 형태로 나타난다. 이런 특징에 바탕해 다음의 규칙 2와 규칙 3을 도출한다.

규칙 2. 뉴스 표제에 나타난 '…, ..., ..'은 모두 말줄임표로 처리한다.
규칙 3. 뉴스 표제 중간에 나타난 말줄임표는 절 구분자로 처리한다.

(3) "가명 '박운'으로 공립중 다녀... 보디가드도 없이 혼자 통학"
(4) 류시원, 허리 디스크 수술 받아..팬들 격려 이어져

(3), (4)에서 말줄임표는 각각 온점 세 개짜리 '...'와 온점이 두 개인 '..'의 다른 꼴로 나타나지만 규칙 2에 근거하여 모두 말줄임표로 분석이 된다. 이어서 규칙 3을 적용하여 이를 절의 구분자로 처리하면 정확한 구문분석이 이루어질 수 있다.

3) 주어의 분석

뉴스 표제는 주격 조사의 실현율이 낮기 때문에[12] 구문분석에서 주어의 분석에 실패할 가능성이 높다. 따라서 뉴스 표제에서의 위치와 품사 정보, 문장 부호 정보, 형태 정보를 활용하여 주어의 분석을 돕는다.

규칙 4. 다음의 문장 부호, 출현 위치, 품사, 형태 등 네 가지 조건 중에 문장 부호 조건과 출현 위치 조건 중 하나와 더불어 품사 또는 형태 조건 하나 이상을 충족시키는 경우 주어로 분석한다.[13]

12) 황은하(2012)에서는 뉴스 표제 말뭉치에 대한 계량적인 연구 결과를 토대로, 뉴스 표제당 체언의 평균 출현빈도가 6개인데 반해 격조사는 표제당 1개 미만으로 나타나, 격조사의 출현 비율이 극히 낮음을 밝혔다.
13) 황은하(2012)에 따르면, 뉴스 표제에 출현하는 반점은 90% 이상이 주어 다음에 사용되어 주로 주격 표지로 기능한다. 나머지 10%의 오분석 가능성을 낮추기 위해 본 연구에서는 출현 위치, 품사 및 형태 유형 등의 조건을 더해 두 개 이상의 조건을 만족시키는 경우에만 주어로 분석되도록 한다.

- 문장 부호 조건: 반점 앞, 또는 큰따옴표 앞
- 출현 위치 조건: 표제의 첫 어절, 또는 큰따옴표 안에서 첫 어절
- 품사 조건: 명사, 사전 등재 고유명사, 미등재 고유명사, '고유명사 + 명사', '명사 + 명사'형식의 구
- 형태 조건: 1음절 축약형, 1음절 한자 표기

위의 출현 위치, 문장 부호, 품사 및 형태 조건 중에 두 가지 이상의 조건을 만족시키는 경우, 주어로 분석한다.

(5) 국토해양위, 경제위기에도 후원금 넘쳤다
(6) "쌍용차, 이달 20일께 부도 위기 몰려 서둘러 법정관리 신청"
(7) 이 대통령 "정부 경제 목표 어려워질 수도"
(8) 경찰 "'장자연 문건' 수사 대상자 12명 외 1명 더 추가"
(9) 檢, 盧전대통령 심야조사 검토

(5)에서 '국토해양위'는 주어로, 표제의 첫 어절이며 반점 앞에 위치하는 두 가지 조건을 만족시킨다. (6)의 경우, '쌍용차'가 주어로, 큰따옴표 안의 첫 어절이면서 반점 앞에 위치하는 두 가지 조건을 만족시킨다. (7)의 주어는 '이 대통령'인데 큰따옴표 앞에 위치하는 출현 위치 조건과 더불어 '고유명사 + 명사'의 구성이라는 품사 조건까지 두 가지 조건을 만족시킨다. (8) 또한 큰따옴표 앞에 나타난 명사, 즉 출현 위치와 품사 조건까지 두 가지 조건을 만족시키는 '경찰'이 주어이다. (9)에서 '檢'은 1음절 한자 표기라는 형태 조건과 표제의 첫 어절이라는 위치 조건 및 반점 앞에 위치하는 문장 부호의 조건을 두루 만족시키며, 따라서 규칙 4에 근거하여 주어로 정확하게 분석될 수 있다.

4) 서술어의 분석

형태소 분석 단계에서 종결형 어미로 끝나는 경우는 일반 텍스트와 같은 구문분석 작업을 진행한다. 또, 비서술성 명사로 끝나는 경우에도 명사구의 구문분석과 같은 알고리즘을 적용하여 구문분석을 진행한다.

자연어처리의 구문분석에서 종결어미는 서술어를 찾아내는 가장 중요한 단서로 활용된다. 그러나 뉴스 표제의 종결형을 분석한 결과, 종결어미로 끝나는 표제의 비율은 23.7%에 불과했다. 나머지는 종결어미가 아닌 다른 품사, 즉 연결어미, 체언, 체언+조사, 부사 등으로 끝을 맺는 것으로 나타났다. 중요한 것은 뉴스 표제에 종결어미가 나타나지 않았다고 하여 곧 서술어가 없는 것은 아니라는 사실이다. 아래에 표제의 끝에 오는 품사 유형별로 일반 텍스트의 서술어 형식으로 패러프레이징(paraphrasing)할 수 있는 규칙을 작성한다. 아울러, 서술어가 생략된 경우 서술어의 복원 가능성과 복원 방법에 대해서도 기술한다.

서술어의 정확한 분석은, 구문분석 정확률의 향상과 더불어 동사가 아니면서 동사 기능을 하는 형태나 어휘를 분석해냄으로써 KC-Tran의 주요 변환 지식인 동사 패턴 지식을 활용할 수 있도록 하여 결과적으로 정확률 향상에 도움을 줄 수 있다.

가) 서술성 명사[14]로 끝나는 표제의 서술어 분석

형태소 분석 결과, 표제의 끝 어절이 서술성 명사로 분석된 경우에 대한 처리 방법이다. 표제 끝에 오는 서술성 명사의 서술어 실현율은 91.3%로[15], 상당히 높은 것으로 나타났다. 다만, 나머지 8.7%의 서술성이 실현되지 않고 일반 명사로 기능하는 경우를 선별해 내는 작업이 우선되어야 한다. 서술성 명사가 서술성이 실현되지 않는 경우는 명사구로, 대부분 관형어의 수식을 받는 구문적인 특징을 보였다. 이때 관형어는 일반 텍스트의 그것과 마찬가지로, 관형사, 용언의 관형사형, 체언에 조사 '의'가 붙은 세 가지 유형이 나타났다.

14) 서술성 명사에 대한 정의는 학자에 따라 차이를 보이지만, 본 연구에서는 형태적·의미적으로 명사의 특징을 보이는 동시에 논항을 취하는 서술어의 기능을 하는 명사를 지칭한다. 한국어 서술성 명사에 관한 앞선 논의에서는 서술성 명사가 구성 성분으로 들어간 명사구의 내적 구조(이현우, 1995, 최경봉 1995, 1996)와 서술성 명사의 통사적 특성(정희정, 1997), 서술성 명사가 이루는 서술성 명사문의 통사적 특성(이병규, 2001), 말뭉치에서 나타난 서술성 명사의 실현 양상(박현아, 2006) 등의 다양한 내용을 다루고 있다. 본 연구에서는 서술성 명사가 뉴스 표제에서 서술어로 기능을 하는지의 여부에 주안점을 두기로 한다.

15) 본 연구에서 나오는 뉴스 표제에 대한 계량적인 데이터는 모두 황은하(2012)에서 인용한 것임을 밝힌다.

규칙 5. 뉴스 표제의 끝 어절이 서술성 명사면서 앞에 관형사, 관형형 어미, 속격조사가 오는 경우는 명사구로 처리한다.

(10) [사진] 참다랑어 올 첫 경매
(11) 美 대북 군사행동 조심스러운 거론
(12) 4년에 한 번 영화 찍는 '올림픽 배우'의 고민

(10)은 '경매'라는 서술성 명사가 관형사 '첫'의 수식을 받아 서술성이 실현되지 않으며 일반 명사로 기능한 경우이다. (11)에서 서술성 명사 '거론'은 관형어 '조심스러운'의 수식을, (12)는 서술성 명사 '고민'이 관형어 "'올림픽 배우'의"의 수식을 받으면서 서술성이 실현되지 않는 일반명사의 쓰임을 보인다.

한편, 표제의 끝에 오면서 서술어로 기능하는 서술성 명사는 '-하다' 동사의 어간으로 보는 것이 타당하다. 따라서 다음과 같은 규칙을 적용한다.

규칙 6. 뉴스 표제가 서술성 명사로 끝나면서 규칙 4에서 규정한 관형어의 수식을 받지 않는 경우, 끝 어절은 '-하다'를 복원하여 서술어로 처리한다.

(13) 북, 오바마 취임식 김계관 파견 타진(→ 타진하다)
(14) 박연차씨 현금뭉치 거래 모두 조사(→ 조사하다)

(13)의 '타진'과 (14)의 '조사'는 서술성 명사이며, 각각 표제 안에서 서술성이 실현되어 실제 서술어로 기능을 한다. '타진'과 '조사'는 문맥에서 관형사나 관형형 어미가 오는 등의 관형어의 수식을 받지 않으며, 규칙 6에 근거해 '-하다' 접미사를 복원하여 서술어로 정확하게 처리될 수 있다.

나) 연결어미(EC)로 끝나는 표제의 서술어 분석

뉴스 표제 문말에 연결어미가 오는 비율은 낮지 않으며, 그 유형도 아주 다양하다. 그러나 표제 문말에 빈도가 가장 높이 사용되는 연결어미는 '-어/아'와 당위성을 나타내는 '-어야/아야' 두 가지이다. 이에 대해 다음과 같이 규칙 7, 규칙 8을 적용하여 정확한 서술어 분석이 가능하도록 한다.

규칙 7. 뉴스 표제의 마지막 형태소가 연결어미 '-어/아'인 경우, '-ㄴ다/는다'로 변환하여 처리한다.

(15) 오바마 취임 앞두고 수감자들 단식투쟁 늘어(→ 는다)

(15)에서는 용언 '늘다'의 어간에 연결어미 '-어/아'가 붙은 '늘어'가 표제의 서술어 기능을 하며, 이때 '늘다'의 종결형인 '는다'로 패러프레이징을 해도 표제의 의미는 바뀌지 않는다.

규칙 8. 뉴스 표제의 마지막 형태소가 연결어미 '-아야/-어야'인 경우에는 보조용언 '하다'를 복원하여 처리한다.

(16) [중앙 시평] 미국이냐, 중국이냐 이분법 벗어나야(→ 벗어나야 한다)

(16)은 규칙 8을 적용하여 보조용언 '하다'를 복원하여 서술어로 처리하게 된다.

다) 부사(MAG)로 끝나는 표제의 서술어 분석

뉴스 표제 종결형 중에 부사로 끝나는 표제의 하위 유형을 살펴보면, 의성의태어가 가장 많고, 다음은 의문 부사 '왜', 기타 부사의 순이다. 의성의태어는 용언 파생접미사 '-하다'가 붙어 용언 파생이 가능한 경우와 용언 파생이 불가능한 경우, 두 가지 유형으로 나뉘는 것으로 나타났다. 따라서 뉴스 표제의 끝에 오는 부사의 하위 유형과 처리 방법은 다음의 규칙으로 정리한다.

규칙 9. 뉴스 표제의 마지막 어절이 상징부사이면서, 이를 어간으로 하는 '-하다' 용언의 파생어가 존재하는 경우에 접미사 '-하다'를 복원하여 처리한다.

(17) "개성공단 폐쇄되나" 뒤숭숭

(17)의 상징부사 ‘뒤숭숭’은 서술어 기능을 하며, 따라서 ‘뒤숭숭’을 어간으로 하는 파생어 ‘뒤숭숭하다’에서 파생접사 ‘-하다’가 생략된 것으로 보아야 한다.

규칙 10. 뉴스 표제의 마지막 어절이 상징부사이면서 해당 부사를 어근으로 하는 파생어가 없는 경우, 부사 패턴 지식을 활용하여 선택제약 관계에 있는, 생략된 서술어를 복원한다.

(18) 한국 ‘글로벌 혁신지수’ 6위로 껑충 (→ 6위로 껑충 뛰다)
(19) 김옥빈, 칸에서 드레스 분실 소동…박쥐팀 발 동동(→ 발 동동 구르다)

(18)에서 상징부사 ‘껑충’은 선택제약을 받는 ‘뛰다’가 생략된 것으로 추정할 수 있다. (19)의 부사 ‘동동’은 동형어로, ‘작은 북을 잇따라 두드리는 소리’, ‘매우 안타깝거나 추워서 발을 자꾸 구르는 모양’ 또는 ‘작은 물체가 떠서 움직이는 모양’을 각각 나타낸다. (20)에서는 선행어 ‘발’을 단서로 부사 패턴 DB에서 ‘동동’과 제약 관계에 있는 ‘구르다’를 복원할 수 있다.

규칙 11. 상징부사가 아닌 일반부사로 표제가 끝난 경우, ‘-하다’, ‘-되다’ 파생이 가능하면, ‘부사+-하다’로 복원하여 처리한다.

(20) WBC 임창용 투구 논란 계속(→ 계속되다)
(21) ‘엔고현상’ 명동은 이미 일본 관광객 가득(→ 가득하다)

(20)과 (21)의 문말에 오는 일반부사 ‘계속’과 ‘가득’은 각각 ‘-되다’, ‘-하다’ 파생이 가능하다. 이들을 파생접사를 복원하여 동사로 분석함으로써, KC-Tran이 이미 보유하고 있는 동사 패턴 DB를 활용할 수 있게 되며, 나아가 정확한 대역어를 선택함으로써 정확한 번역문을 출력할 수 있게 된다.

라) 어근으로 끝나는 표제의 서술어 분석

뉴스 표제의 끝에 오는 서술어는 용언 파생접미사가 생략된 채 어근으로

끝나는 경우가 적지 않다. 이 경우, '-하다'를 복원하여 서술어로 처리한다.

규칙 12. 뉴스 표제의 끝에 어근이 오는 경우, '-하다' 등의 접사를 복원하여 서술어로 처리한다.

(22) 2009년 北상황 1990년대보다 심각(→ 심각하다)

(22)에서 어근 '심각'은 '심각하다'로 용언의 형태를 복원하여 처리함으로써 정확한 구문분석이 가능하게 되며, 나아가 KC-Tran이 기존에 보유한 동사 패턴 지식을 활용할 수 있게 된다.

마) 의존명사 '듯'으로 끝나는 표제의 서술어 분석

뉴스 표제의 끝에 의존명사 '듯'으로 끝나는 경우가 적지 않은데, 보조형용사 '듯하다'에서 접사 '하다'가 생략된 꼴로 보는 것이 옳다. 따라서 다음과 같은 규칙을 적용하여 서술어 분석이 가능하도록 한다.

규칙 13. 뉴스 표제의 끝에 의존명사 '듯'이 오는 경우, 접사 '-하다'를 복원하여 서술어로 처리한다.

(23) 노인 시중드는 '실버 로봇' 5년 내 등장할 듯
→ 노인 시중드는 '실버 로봇' 5년 내 등장할 듯하다
(24) 후계문제 둘러싼 비상사태 대비책인듯
→ 후계문제 둘러싼 비상사태 대비책인 듯하다

(23), (24)는 모두 의존명사 '듯'으로 끝난 표제로, 규칙 13을 적용하여 각각 '-하다'를 복원해 줌으로써 서술어의 정확한 분석이 가능하게 된다.

3.2.2 한중 변환 모듈

1) 한자 표기

뉴스 표제에 나타난 한자 표기는 괄호(()) 안에 써서 선행 한글 어절의 한자 표기를 제시하는 경우와 괄호 없이 한글 대신 쓰는 두 가지의 경우가 있다. 이를 규칙으로 정리하면 각각 규칙 14, 15와 같다.

규칙 14. 괄호 안에 나타난 한자 표기는 괄호의 선행 어절(이때 어절의 길이는 괄호 안의 한자의 글자수와 같음)의 한자 표기이므로, 분석 모듈에서 선행 어절의 변환 근거로 사용한다.

(25) 노(盧) 전(前)대통령측, 사저 출발모습 언론에 공개키로
(26) [조선데스크] 구글과 한국 정부의 일전(一戰)

규칙 14에 근거하여 (25)의 '노'는 '盧'로, (26)의 '일전'은 '一戰'으로 간단하게 변환할 수 있다.

규칙 15. 괄호가 없이 쓰인 한자 표기는 한글로 변환한 후 한국어 형태소 분석을 하되, 띄어쓰기가 있든 없든 뒤에 오는 어절과는 무관한 단독 어형으로 분석한다.

(27) 美오클랜드서 백인 경관이 흑인청년 사살
(28) 中 '글로벌 언론 공정' 나섰다

규칙 15에 근거해, (27)의 '美'는 '미'로 변환한 후, 1음절 고빈도 어휘 사전에 근거하여 '고유명사'로 처리할 수 있다. (28)도 마찬가지로 '中'을 '중'으로 변환한 후 1음절 고빈도 어휘 사전에 근거하여 '고유명사'로 처리하게 된다. 괄호가 없이 쓰인 한자 표기는 뒤에 띄어쓰기를 하지 않는 경우가 많으나, 형태소 분석을 할 때 반드시 뒤의 한글로 시작되는 형태소와는 분리하여 분석해야 정

확한 결과를 도출할 수 있다.

2) 줄임표

황은하(2013)에 따르면, 중국어는 한국어와 달리 띄어쓰기가 없지만, 뉴스 표제에서는 띄어쓰기를 절 구분자로 사용한다. 따라서 앞서 규칙 3을 이용하여 절 구분을 정확히 한 뉴스 표제는 분석 모듈에서 적당한 대응 형식을 제시해 줄 필요가 있다.

규칙 16. 한국어의 절 구분자 줄임표는 중국어에서 띄어쓰기로 변환한다.

이를 적용한 예를 보이면 다음과 같다.

(29) 말로만 '속도전' … 뜯어보니 '지구전'
(→'就业机会本属速度战 政府却暗操持久战)

3) 반점의 변환

중국어 뉴스 표제에서 반점은 아주 제한적으로 사용된다[16]. 한국어의 주어 구분자로 기능하는 반점은 중국어에서 표기할 필요가 없다. 따라서 반점의 변환에 대해 다음과 같은 규칙을 작성하여 적용한다.

규칙 17. 주어 뒤에 오는 반점(,)은 변환 과정에서 제거한다.[17]

16) 중국어 일반 텍스트에서 반점은 주로 1) 문장 안에서 주어와 서술어 사이에 휴지를 둘 때, 2) 동사와 목적어 사이에 휴지를 둘 때, 3) 부사어 뒤에 휴지를 둘 때, 4) 복문에서 절과 절 사이에 휴지를 둘 때 사용한다. 그러나 황은하(2012)에 근거하면, 뉴스 표제에서는 상기의 세 가지 기능 중 1)의 경우에만 사용되는 것으로 관찰되었으며, 출현빈도는 전체 문장 부호 중 3.01%로 나타나, 한국어 뉴스 표제에 사용된 반점(18.56%)보다 훨씬 제한적으로 사용되는 것으로 나타났다.

17) 여기서 제거되는 반점은 앞선 분석 단계에서 이미 주어의 분석 규칙에 활용된 것으로, 잉여적인 것이다.

(30) 미(美), 연비 16㎞/L 안 되는 승용차 2016년부터 판매 금지
(→'美国2016年起将禁止出售低燃油效率汽车)

(30)의 주어 '미' 뒤에서 주어 구분자로 기능하는 반점(,)은 규칙 17을 적용하여 중국어 변환 과정에 삭제되며, 따라서 정확한 번역문을 출력할 수 있게 된다.

4) 작은따옴표('')의 변환

규칙 18. 한국어의 작은따옴표는 중국어의 큰따옴표로 변환한다.

(31) 'MC몽의 연인' 주아민, 비키니 몸매 공개했다 삭제 (→"MC梦恋人"朱雅敏公开比基尼照片)

(31)에서 작은따옴표는 규칙 18을 적용하여 중국어 번역문에서 큰따옴표로 표시되도록 한다. 이 경우, 작은따옴표로 출력을 해도 번역문의 의미 변화에 영향을 미치지는 않지만, 중국어 뉴스 표제의 문장부호 사용법에 걸맞는 번역문을 출력해 준다는 데 의미가 있다.

5) 미등재 고유명사의 변환

한국어 뉴스 표제에 나타난 미등재 고유명사의 대부분은 외국인 인명 등의 고유명사라는 점을 감안하여 다음과 같은 규칙을 작성한다.

규칙 19. 규칙 1에서 미등재 고유명사로 분석된 형태소는 로마자화(romanizaiton)하여 출력한다.

(32) 미네르바, 대선 때 MB 퇴진 운동 단체 회원 가입해 활동(→ 'Minerba)

3.2.3. 중국어 생성 모듈

한중 변환 모듈을 통해 중국어로 번역된 결과물을 보다 '중국어답게' 다듬어 주는 데 필요한 지식은 중국어 뉴스 표제의 특징을 토대로 작성되어야 한다. 이를 기술해 보이면 다음과 같다.

규칙 20. 인용문 앞에 나타나는 주어, 즉 발화 주체가 나타나는 경우, 중국어에서는 주어 뒤에 쌍점(:)을 더하고, 직접인용문은 문장 부호가 없이 기술한다.

(33) 青 "비상경제상황실 설치"
(→青瓦台：設立非常經濟情況室(청와대: 비상경제상황실 설립))

(33)에서 '青'은 직접인용문의 발화 주체로 규칙 4에 근거하여 주어로 분석된다. 이어서 규칙 20을 적용하여 중국어 문장부호 사용 격식에 맞는 출력문으로 다듬어 준다.

이상의 20가지 규칙 외에도 한중 대조 분석 연구 결과를 토대로 기계번역에 긍정적인 효과를 가져다 줄 다음과 같은 언어 데이터베이스, 또는 규칙을 작성하여 한국어 분석 모듈에서 활용할 수 있을 것이다.

첫째, 한국어 뉴스 표제에는 1음절 형태·어휘의 출현 비율이 상당히 높다. 뉴스 표제에 나타난 형태·어휘 중 1음절 비율은 일반 텍스트보다 2배 이상 높은 것으로 확인되었다. 따라서 형태소 분석기의 구현 방식에 최단일치법을 도입하는 것을 고려해 볼 수 있다.

둘째, 미등재어 형태·어휘 유형의 목록과 빈도 정보는 뉴스 표제의 한국어 분석에서 긍정적인 효과를 낳을 수 있을 것이다.

셋째, 뉴스 표제의 조사 실현율은 아주 낮은데, 그 중에 부사격조사와 보조사의 실현율이 상대적으로 높게 나타났다. 부사격조사에서 장소를 나타내는 '-에서'의 실현율이 가장 높으나, 주로 준꼴인 '-서'로 실현된다. 또, 보조사 '-은/는'은 많은 경우 '-ㄴ'으로 실현된다. 따라서 형태소 분석 단계에서 '-에서'의 준꼴인 '-서'와 '-은/는'의 준꼴 '-ㄴ'에 대해 가중치를 부여하는 방법을 권장한다.

4. 기계번역 적용 및 평가

기계번역 평가(MT evaluation), 즉 번역의 정확성, 번역 속도 등의 기계번역 시스템의 성능에 대한 체계적이고 객관적인 평가는 개발자에게는 시스템의 문제점을 발견하도록 돕고 개선 방향을 제시해 주며, 사용자에게는 시스템에 대한 올바른 이해를 가능케 한다. 아놀드 외(Arnold et al, 1993)[18]는 기계번역 평가 방법을 연산 평가(Operational evaluation)[19], 선언적 평가(Declarative evaluation), 유형적 평가(Typological evaluation), 유형적 비교평가(Typological comparative evaluation)[20] 등의 네 가지로 소개하였다.

본 연구의 실험 평가는 3장에서 제시한 20가지의 규칙을 기계번역기에 적용하여, 그 전후 출력문이 어떻게 달라졌는지를 두고 진행할 것이다. 우선, KC-Tran의 일반 텍스트의 번역 정확률과 비교하기 위해 한국전자통신연구원에서 사용하는 선언적 평가 방법을 채택한다. 다음으로, 본 연구의 대조 분석 결과를 바탕으로 정리한 뉴스 표제 지식이 기계번역의 정확률에 얼마나 영향을 미치는지를 측정하기 위해 유형적 평가를 수행하고자 한다.

4.1 선언적 평가 방법

번역에서는 원저자의 의도를 충분히 살려 원문의 의미를 정확하게 전달하는 것이 무엇보다 중요하다고 해야 할 것이다. 따라서 번역의 정확성은 원문과 번역문이 의미적으로 얼마나 가까운가 하는 번역의 등가성(translation equivalence)

---

18) 이민행 외(1998)에서 재인용.

19) 연산 평가는 사용자 중심의 결과 중심의 종합적 평가로, 경제성 분석에 주안점을 두며, 기계번역의 단어당 비용(cost-per-word)과 후편집(post-editing)에 드는 비용을 계산한다. 유형적 평가는 테스트의 도구가 되는 측정 장치(test suit)를 사용하여 평가하며, 시스템 개발자에게는 어떤 구조가 문제점을 갖는지를 분명하게 알 수 있게 함으로써 똑같은 측정장치를 다른 시스템에도 적용하여 비교할 수 있도록 하는 장점이 있다.

20) 유형적 비교평가는 3단계 평가 방법을 채택하여, 1단계에서는 측정장치의 각 평가 문항에 포함된 전문인 번역을 기계에 의한 번역결과와 비교하고, 2단계에서는 그 결과 자체를 여러 평가기준에 의해 일차적으로 개별 평가하며, 3단계에서 그 개별 평가 결과를 계량화하여 종합적으로 평가하는 방법이다.

에 의해 평가된다. 선언적 평가는 바로 이런 관점에서 출발한 평가 방법으로, 입력과 출력 모두에 명료성(Intelligibility), 정확성(Accuracy), 문체(Style) 등을 기준으로 평가한다. 여기서 명료성은 번역된 문장이 목표언어(target language)로 잘 되어 있는가를 지수로 표현하고, 정확성은 원천언어로 된 텍스트(source text)의 내용이 얼마만큼 잘 보존되어 있는가를 지수로 나타낸다(이민행 외, 1998).

KC-Tran은 기존의 평가에서 바로 이와 같은 선언적 평가 방법을 적용하고 있으며, 주로 명료성과 정확성에 주안점을 두어 평가해 왔다. 평가 결과는 정확률로 계산하여 보여주며, 평가 기준은 아래와 같이 4점부터 0점의 다섯 단계의 점수를 채택하며 점수 뒤에 '-' 표지를 더하여 보다 엄밀한 평점을 시도한다.

〈표 2. 기계번역의 선언적 평가 기준〉

점수	설명
4	원어의 의미가 그대로 전달된 경우(meaning preserved)
3	중국어 표현이 어색하고, 원문의 의미가 부분적으로 전달된 경우 (meaning partially preserved)
2	문장 중에 단 하나의 구라도 정확히 번역된 경우 (phrase level translated)
1	문장 중에 단 하나의 단어라도 정확히 번역된 경우 (word level translated)
0	번역문 출력이 안 된 경우(no output)
-	위의 1~4점에서 두 점수 가운데 하나를 주기 애매한 경우

이와 같은 선언적 평가 방법은 한국어 입력문에 대한 정확한 이해와 더불어 중국어 출력문의 번역의 질에 대한 평가가 이루어져야 하므로 한국어와 중국어의 이중언어(bilingual) 화자에게 평가를 의뢰한다. 또한, 정확성과 명확성에 대한 판단에 있어 주관성을 철저히 배제하기는 어렵다고 보아, 두 명의 이중 언어 화자에게 평가를 의뢰하여 평균 점수를 산출함으로써 가능한 한 객관적인 평가 결과를 얻을 수 있도록 한다. 또한, 선언적 평가 시 번역문을 제시하는 것이 일반적이지만, 본 연구에서는 언론사의 한중 뉴스 표제 번역문이 치환, 생략, 삽입, 변조 등의 다양한 번역 전략을 구사함으로써 원문에 대한 충실도 또한 높지 않기 때문에 평가자에게 번역문을 제시하지 않았다. 그보다는 숙련된 한국어와 중국어 실력을 갖춘 이중언어 화자의 언어적 직관에 따라 평가하도록 한다.

4.2 평가 결과 분석

선언적 평가 세트는 뉴스 표제 말뭉치에서 언론사를 구분하지 않고 표제 100개를 무작위로 추출하여 사용한다. 평가용 표제를 살펴보면 어절 수 및 표제 당 평균 어절 수는 다음과 같다.

〈표 3. 평가용 표제 세트의 개요〉

항목	합계
문장 수(문장)	100
총 어절 수(어절)	612
표제 평균 길이(어절)	6.1

뉴스 표제의 평균 길이는 6.1어절로 한국어 뉴스 표제의 평균 길이인 5.9어절과 비슷하다.[21)]

4.2.1 정확률 향상에 대한 평가

상기의 100개의 표제를 입력문으로 삼아, KC-Tran 기계번역 시스템에 3장에서 기술한 언어학적 지식을 적용하기 전과 후로 나누어 각각 번역문을 출력하고 평가하여 다음과 같은 결과를 얻었다.

〈표 4. KC-Tran의 뉴스 표제 기계번역 평가〉

평점	지식 적용 전		지식 적용 후	
	표제수	점유율(%)	표제수	점유율(%)
4 ~ 4-	16	16.0%	26	26.0%
3 ~ 3-	37	37.0%	42	42.0%
2 ~ 2-	33	33.0%	26	6.0%
1 ~ 1-	13	13.0%	6	6.0%
0	1	1.0%	0	0.0%
계	100	100%	100	100%

21) 황은하(2012)에서 한중 뉴스 표제 병렬말뭉치를 소개하면서 산출한 수치로, 이를 인용한 것이다.

표에서 보이는 것과 같이, 0~2점 구간의 점수를 받은 결과물의 수가 3장의 언어 지식 적용 전의 45%에서 적용 후에는 12%로 확연히 줄었다. 이와 동시에 3~4점 구간의 높은 점수를 받은 결과물의 수는 지식 적용 전의 53%에서 지식 적용 후 68%로 적잖게 늘었다.

위에서 얻은 점수를 다음의 정확률 계산 수식에 대입하여 계산하면 백분율로 표시되는 정확률을 얻을 수 있다.

정확률 = (총점수 / 만점) X 100

이상과 같은 평가를 통해 지식 적용 전후의 KC-Tran의 정확률은 다음과 같다.

지식 적용 전 정확률 = (241.5/ 4 X 100) X 100 = 60.4%
지식 적용 후 정확률 = (274.5 / 4 X 100) X 100 = 68.8%

지식 적용 전의 뉴스 표제의 기계번역 정확률은 60.4%로, 한중 기술 문서 자동번역 기술의 정확률 80%, 한중 방송 자막 자동 번역 시스템 셋탑박스 탑재 기술의 뉴스 정확률 80%보다 각각 약 20% 낮은 수치다.[22] 언어 지식 적용 후에는 68.8%의 정확률을 얻어 지식 적용 전에 비해 정확률이 8.4% 향상, 비교적 크게 개선된 것을 알 수 있다.

4.2.2 언어 지식 기여도에 대한 평가

100개의 평가 대상 표제에서 59개가 2장에서 정리한 규칙을 적용하여 패러

22) 박상규(2009)에 따르면 한국전자통신연구원(ETRI)의 다국어 기계번역 기술 개발 수준은 다음과 같다.
- 영한 특허문서 자동번역 기술: 정확률 85%
- 한영/영한 기술논문 자동번역 기술: 한영 정확률 75%, 영한 80%
- 한중 기술문서 자동번역 기술: 정확률 80%
- 영한/중한 웹신문 자동번역 기술: 영한 정확률 75%, 중한 80%
- 한중 방송자막 자동번역 시스템 및 셋탑박스 탑재 기술 개발(뉴스 정확률 80%, 드라마 75%)

프레이징에 성공했고, 규칙의 적용 횟수는 총 79회인 것으로 나타났다. 규칙별 적용 빈도를 보이면 다음과 같다.

〈표 5. 뉴스 표제 기계번역 규칙 의존도〉

규칙 번호	규칙 설명	적용 빈도	적용 비율
2, 3	절 분석	9	11.4%
4	주어 분석	16	20.3%
5	서술성 명사의 서술어 분석	23	29.1%
6	연결어미 '-아/어' 종결형 표제의 서술어 분석	4	5.1%
7	연결어미 '-아야/어야' 종결형 표제의 서술어 분석	3	3.8%
13	의존명사 '듯' 종결형 표제의 서술어 분석	5	6.3%
14	괄호 안의 한자 표기의 활용	2	2.5%
15	괄호가 없는 한자 표기의 활용	1	1.3%
20	발화 주체가 있는 직접 인용문의 문장 부호	12	15.2%
명사 패턴DB	일반명사로 끝나는 뉴스 표제의 서술어 분석	4	5.1%
합계		79	100.0%

뉴스 표제 기계번역 규칙 의존도를 살펴보면 다음과 같은 몇 가지 특징을 보인다.

첫째, 모두 20개의 규칙 중에 2, 3, 4, 5, 6, 7, 13, 14, 15, 20번까지 10개의 규칙과 일반명사로 끝나는 표제의 서술어 분석을 위한 명사 패턴 DB가 패러프레이징에 성공적으로 응용된 것을 알 수 있다. 이 규칙들이 뉴스 표제의 보다 일반적인 특징을 다루고 있기 때문인 것으로 해석된다.

둘째, 규칙 5, 6, 7, 13은 모두 뉴스 표제의 서술어 분석을 위한 규칙으로, 전체 규칙 적용 횟수의 44.3%를 차지한다. 규칙 5(29.1%)가 가장 많이 적용되었는데, 이는 서술성 명사가 뉴스표제의 끝에서 서술어로 기능하는 경우가 그만큼 일반적이기 때문이다. 여기서 서술어 분석 규칙은 서술어를 정확하게 가릴 수 있도록 도울 뿐만 아니라, 결과적으로 기존의 동사 패턴 DB를 활용할 수 있도록 하기 때문에 정확률 향상에 큰 도움을 준다.

셋째, 규칙 4, 즉 주어 분석에 관한 규칙의 적용률은 20.3%로, 2위를 차지

한다. 이 규칙은 정확한 주어의 분석을 가능케 해 줌으로써 동사 패턴 DB에서 정확한 동사 패턴을 선택할 수 있는 근거를 더하는 역할을 한다.

넷째, 규칙 20의 적용률도 15.2%로, 결코 낮지 않은 의존도를 보인다. 규칙 20은 생성을 위한 규칙으로, 정확률을 높이는 데 직접적으로 기여하지는 않지만, 중국어 뉴스 표제 문법에 알맞은 출력문을 생성하는 데 기여한다.

다섯째, 규칙 2와 3, 즉 절의 분석 규칙에 대한 의존도는 11.4%로, 네 번째로 높은 것으로 나타났다. 이 규칙은 절 구분을 정확히 함으로써 다음 단계의 구문분석의 정확률을 높이는 데 기여하며, 궁극적으로 번역의 정확률 향상에 도움을 준다.

여섯째, 20개 규칙 중 나머지 10개의 규칙은 평가 세트에 적용된 예가 없었다. 이는 해당 규칙들이 뉴스 표제의 언어 특징 중 덜 보편적인 특징을 다루고 있기 때문인 것으로 풀이된다. 뉴스 표제 평가 세트의 규모를 늘리고 입력문의 유형이 다양해지면 적용 가능한 규칙의 유형도 늘어날 것으로 예상된다.

이밖에 규칙이 적용되지 않은 나머지 52%에 대해 관찰하여 규칙이 적용되지 않은 이유를 살펴보면 뉴스 표제 입력문이 명사구 또는 완전 종결형으로 되어 있어 규칙 적용이 필요하지 않았다. 한편, 규칙이 적용된 59개 표제 중에 37개의 표제가 지식 적용 후 평가 점수가 높아졌고, 17개는 지식 적용 전과 평가 점수가 같았으며, 5개는 지식 적용 후 점수가 낮아진 것으로 나타났다. 지식 적용 후 점수가 낮아진 예들을 살펴보면 그 원인은 다음과 같다.

(34) 서울대 육상부, 성적은 꼴찌-열정은 금메달
→ 규칙 4 적용 후: 서울대 육상부가 성적은 꼴찌-열정은 금메달
(35) “장자연 소속사 전 대표 신병 이번 주 인도 요청”
→ 지식 적용 전 번역문: 葬自然所屬公司全代表新兵本周人道請求
→ 규칙 5 적용 후: “장자연 소속사 전 대표 신병 이번 주 인도 요청하다”
→ 규칙 적용 후 번역문: 要求"庄子延所屬公司全代表新兵本周人道

(34)는 규칙 4가 적용된 후, 뉴스 표제가 이중주어문의 형식을 띠게 되어

구문분석에서 난이도가 높아진 탓에 번역문의 정확률이 오히려 낮아진 것으로 분석된다. (35)는 조사가 하나도 실현되지 않아 구문분석의 난이도가 높은 예로, 지식 적용 전에는 구문분석에 실패하면서 한국어와 같은 어순으로 배열되는 결과(2점)를 만든 반면에, 지식 적용 후 '요청하다'의 논항을 "장자연 소속사 전 대표 신병 이번 주 인도"로 잘못 분석하면서 번역 결과가 오히려 원래보다 낮은 평점 '2-점'을 받은 경우이다.

이상으로, 선언적 평가 방법을 통해 뉴스 표제의 기계번역 정확률이 지식 적용 전의 60.4%에서 지식 적용 후 68.8%로, 8.4% 향상된 것을 확인하였다. 이를 통해 기계번역에 있어서, 언어간 대조 분석을 통한 언어학적 지식 및 그에 기반하여 작성된 번역 규칙의 기여도가 일부 입증되었다.

5. 결론

본 연구는 1장에서 연구의 목적을 기술하고 연구의 필요성 및 논의의 구성에 대해 기술하였다. 2장에서는 한중 기계번역의 국내외 기술 동향에 대한 개요와 더불어 실험 적용 대상인 KC-Tran의 시스템 구성, 즉 한국어 분석, 한중 변환 및 중국어 생성부 등의 3개 모듈로 나누어 모듈별로 적용 가능한 한중 대조 분석을 통한 언어학적 지식을 20개의 규칙으로 정리하였다. 3장에서는 2장에서 정리한 규칙을 패러프레이징 기법을 이용해 입력문을 변형하여 평가한 결과, 지식 적용 후 정확률이 8.4% 향상된 것을 확인할 수 있었다. 또한, 2장에서 정리한 20개에 대한 입력문의 의존도를 살펴봄으로써, 기계번역을 위해 정리한 규칙의 기여도를 객관적으로 입증하였다.

본 연구는 한중 뉴스 표제가 일반 텍스트의 그것과는 구분되는 언어적 특징을 규칙으로 정리하여 기계번역에 응용하고 평가함으로써, 기계번역에 있어서 언어학적 지식의 필요성과 기여도를 재차 입증하였다. 본 연구의 의의는 두 가지 측면에서 평가될 수 있다. 우선, 연구 대상을 한중 뉴스 표제로 한정하긴 했지만, 적은 수의 규칙으로 8.4%의 정확률 향상에 기여했다는 데서 실천적 의미를 찾을 수 있다. 다음으로, 90년대만 해도 언어학적 지식에 기반한 기계번역

기법이 주류를 이루었던 데 반해, 최근 빅데이터와 통계에 기반한 기법이 새로운 추세를 이루는 이 시점에 언어학적 지식의 기계번역에 대한 기여도를 재차 입증했다는 데 또 다른 의의가 있다.

다만, 이와 같은 언어 지식에 기반한 연구를 뉴스 표제뿐만 아닌 다양한 장르의 텍스트에 확대 적용하는 일은 앞으로 연구할 또 다른 과제로 남긴다.

참고문헌

구글『번역이 이루어지는 과정 알아보기』, http://translate.google.co.kr/about/intl/ko_ALL

박상규 (2009)『응용특화 한중영 자동번역 기술 개발에 관한 연구』, 한국전자통신연구원.

박현아 (2006)『한국어 서술성 명사의 실현 양상 연구』, 고려대학교 대학원 석사학위논문.

서상규・한영균 (1999)『국어정보학 입문』, 태학사.

서상규 (2009)「국어 특수 자료 구축의 성과와 전망」,『새국어생활』19(1): 35-57.

송경화 (2006)『신문 기사의 코퍼스 언어학적 분석』, 고려대학교 대학원 석사학위논문.

안동환 역 (2008)『코퍼스기반 번역학: 이론, 연구결과, 응용』, 도서출판 동인.

유현경・황은하 (2009)「병렬말뭉치 구축과 응용」,『언어사실과 관점』25: 5-40,

이민행 (1998)「독-한 명사구 기계번역시스템의 구축」,『언어와 정보』2(1): 79-105.

이민행 외 (1998)「기계번역 시스템 측정 장치 연구」,『언어와 정보』2(2): 185-220.

이병규 (2001)『국어 술어명사문 연구』, 연세대학교 대학원 박사학위논문.

이봉원 (2002)「북한의 언어공학 현황과 발전 전망에 대한 연구」『(2001 신진연구자) 북한 및 통일 관련 논문집(제5권): 북한실태(사회)』, 통일부.

이인묵 (2012)「구글 기계번역 책임자 베누고팔」,『조선비즈』2012년 11월 3일, http://m.biz.chosun.com/svc/article.html?contid=2012110201300.

이현우 (1995)『현대 국어의 명사구의 구조 연구』, 서울대학교 대학원 박사학

위논문.

정희정 (1997)「서술성 명사의 통사적 특성」,『언어 정보와 사전 편찬』7: 7-34.

최경봉 (1996)『국어명사의 의미 구조 연구』, 고려대학교 대학원 박사학위논문.

황은하・홍문표・최승권 (2002)「동사 패턴에 기반한 한중 기계번역」,『한국중국언어학회 2002년 춘계 학술대회 논문 발표집』, 한국중국언어학회.

황은하 (2009)「한중 인터넷 뉴스 표제 병렬말뭉치 연구: 기계번역을 위한 시험적 연구」,『번역학연구』10(3): 217-245.

황은하 (2012)『한중 뉴스 표제의 대조 분석 및 기계번역 응용』, 연세대학교 대학원 박사학위논문.

황은하 (2013)「말뭉치에 기반한 한중 뉴스표제의 문장부호 번역 연구」,『번역학연구』14(2): 283-311.

尹世超 (2001) ≪標題語法≫, 商務印書館

Arnold, D., Sadler, L. and Humprheys, R.-L. (1993) *Evaluation: An Assessment*. Machine Translation 8, 1-24.

Diganta, S., Sivaji, B. (2005) *A Semantics-based English-Bengali EBMT System for translating News Headlines*. In Workshop on Example-Based Machine Translation. MT SUMMIT X.

John Hutchins (2005) *Example-based machine translation - a review and commentary*. Vol. 19, No. 3/4.

Kenji Ono (2003) *"Translation of News Headlines"*, in Proceedings of MT Summit IX.

R.M.K. Sinha (2002) *Translating News Headings from English to Hindi.* in Proceeding (357) Artificial Intelligence and Soft Computing.

Takehiko Yoshimi (2001) *Improvement of Translation Quality of English Newspaper Headlines by Automatic Pre-editing.* Machine Translation Volume 16, Number 4.

[Abstract]

Linguistic Knowledge-Based Korean-Chinese Machine Translation of News Headlines

Huang, Yinxia
(Paichai University)

The aim of this study was to conduct a contrastive analysis on the linguistic characteristics of Korean and Chinese news headlines and to apply the research results to machine translations of Korean and Chinese news headlines. To achieve this purpose, this study constructed a parallel corpus and comparable corpus of Korean and Chinese news headlines to conduct a comparative analysis and correspondence study regarding the linguistic characteristics presented in the form, vocabulary, syntax and punctuation marks of Korean and Chinese news headlines. The results from the contrastive analysis were applied to the machine translation of Korean and Chinese news headlines to verify the usefulness of this study.

The significance of this study can be largely observed in linguistic and language engineering aspects. First, this study described the linguistic use of Korean news headlines that can be differentiated from general text. In addition, this study clarified the similarities and differences between Korean and Chinese news headlines, and summarized the correspondence rules required in Chinese translation. Next, by applying the experimental results and evaluating the linguistic research results in a language engineering study via machine translation, this study not only enhanced the accuracy rate, but also presented a successful case of interdisciplinary integration between linguistics and language engineering.

▸ Key Words: Korean-Chinese machine translation, linguistic knowledge-based machine translation, news headlines, machine translation assessment, declarative evaluation

황은하(黃銀霞)
배재대학교 한국어문학과
behisson@gmail.com
관심분야: 기계번역, 말뭉치언어학, 대조언어학, 사전편찬학

논문투고일: 2014년 10월 31일
심사완료일: 2014년 11월 25일
게재확정일: 2014년 12월 4일