

## 특허 기계 번역에 대한

# RIBES 한국어 자동평가 문제에 대한 고찰

장현진<sup>○</sup>, 장문석, 노한성

한국특허정보원

(namuland, cms1004, neodream)@kipi.or.kr

## KoRIBES : A Study on the Problems of RIBES in Automatic Evaluation

### English-Korean Patent Machine Translation

Hyeon-Jin, Jang<sup>○</sup>, Moon-Seok, Jang, Han-Sung Noh

Korea Institute of Patent Information

#### 요 약

자연어 처리에서 기계번역은 가장 많이 사용되고 빠르게 발전하고 있다. 기계번역에 있어서 사람의 평가가 가장 정확하고 중요하지만 많은 시간과 비용이 발생된다. 이에 기계번역을 자동 평가하는 방법들이 많이 제안되어 사용되고 있지만, 한국어 특성을 잘 반영한 자동평가 방법은 연구되지 않고 있다. BLEU와 같은 자동평가 방법을 많이 사용하고 있지만 언어의 특성 차이로 인해 원하는 평가결과를 얻지 못하는 경우가 발생하며, 특히 특허나 논문과 같은 기술문서의 번역에서는 더 많이 발생한다. 이에 본 논문에서는 단어의 정밀도와 어순이 평가에 영향이 있는 RIBES를 가지고 특허 기계 번역에서 영어→한국어로 기계 번역된 결과물의 자동평가에 대해 사람의 평가와 유사한 결과를 얻기 위해 tokenization 과정에서 복합 형태소 분리를 통한 평가방법을 제안하고자 한다.

주제어: 기계 번역(machine translations), 특허 번역(patent translation), 자동 평가(automatic evaluation), BLEU, RIBES

#### 1. 서론

인공지능의 발전으로 사람이 해야 하는 많은 부분이 기계가 대신하고 있고, 기계번역 시스템도 인공지능의 발전에 선두에서 많은 발전을 이루었다. 다양한 언어를 사용하는 사람들 사이에서 언어를 공부하지 않아도 전 세계 사람들과 의사소통이 가능한 시대가 오고 있는 것이다. 기계번역 기술은 과거에도 존재하였지만, 문법 규칙 또는 통계 기반으로만 번역을 했기 때문에 정밀도가 낮았다. 딥러닝 기술의 발전으로 통계나 규칙 기반이 아닌 대용량 학습데이터를 사용한 인공신경망 기계번역이 좋은 성능을 보이고 있고 대표적인 기계 번역엔진인 구글(Google)의 번역 서비스는 전 세계에서 매일 평균 1,000억 건 이상 번역되고 있을 만큼 많이 사용되고 있다. 구글이나 아마존과 같은 일반적으로 사용되는 기계번역이 보편화되기 이전부터 특허에서는 기계번역이 활발하게 사용되어 왔으며, 유럽연합(EPO), 세계지적재산권기구(WIPO, World Intellectual Property Organization) 등에서 자체적으로 기계 번역기를 구축하여 해외 특허정보 이용자들이 관심 있는 특허정보를 번역하여 쉽게 파악 할 수 있도록 서비스하고 있다[1].

이와 같이 기계번역은 널리 사용되고 있어서 번역품질을 향상하기 위해 번역 결과에 대한 평가는 매우 중요하다. 번역평가는 번역된 문장이나 문서를 사람이 이해하는 정도를 표현하는 것이기 때문에 사람이 직접 번역문을 읽고 이해정도를 평가하는 것이 가장 정확하고 신뢰도가 높지만, 기계번역의 많은 번역문을 직접 평가하기

란 엄청난 시간과 막대한 비용이 발생되기 때문에 BLEU(Bilingual Evaluation Understudy)[2]와 같은 자동평가 방법이 연구되고 있다.

하지만 한국어의 특성을 잘 반영한 자동평가 방법은 거의 연구되지 않고 있다. 일반적인 자동평가 방법을 사용하여 번역 결과를 평가하면 언어의 특성 차이로 인해 원하는 평가결과를 도출할 수 없는 경우가 발생하며, 특히 같은 기술문서는 더욱이 잘 사용하지 않는 기술용어나 합성어, 외래어의 한글표현의 다양성, 서술어의 축소 등으로 인해 기계번역뿐만 아니라 사람이 하는 번역과 평가도 매우 까다롭다. 특허에서 번역은 불완전하거나, 오역 등의 오류로 인해 분쟁을 일으킬 수 있고 번역의 수요자와 공급자간의 권리 의무 관계를 명확히 설정할 수 있도록 해주는 역할을 해야 하기 때문에 여러 나라에서 출원·등록된 특허의 번역이 잘되어서 명확하게 의미를 전달 할 수 있어야 할 것이다[3].

이에 본 연구에서는 특허 기계번역에서 영어→한국어로 기계 번역된 결과물의 자동평가를 위해 자동평가 방법 중 단어와 단어사이의 n-gram으로 평가하는 BLEU의 특징을 소개하고 단어의 정밀도와 어순을 평가하는 RIBES(Rank-based Intuitive Bilingual Evaluation Score)[4]를 활용하여 특허 기계번역의 한국어 번역 평가 시 원하는 결과도출을 위한 방법을 제시함으로써 특허 기계번역에서 한국어 자동 평가방법을 좀 더 신뢰할 수 있도록 하고, 기계번역 결과물의 품질을 높이는데 더 적극적으로 활용될 수 있도록 기반을 마련하고자 한다.

## 2. 관련 연구

### 2.1. BLEU 자동평가 방법

BLEU는 가장 많이 사용되는 자동평가방법으로 단어끼리의 n-gram의 비교를 통해 기계번역문과 사람이 번역한 문장 사이의 겹치는 부분을 측정하여 점수로 산출한다.

$$BLEU = BP \times (p_1 p_2 p_3 p_4)^{\frac{1}{4}}$$

$$BP(Brevity Penalty) = \min(1, \exp(1 - \frac{r}{h}))$$

그러기 위해 BLEU는 문장을 단어단위로 나누어 n-gram을 평가하게 되는데, 문장(Sentence)을 띄어쓰기 기반으로 하여 단어를 구분하기 때문에 한국어에서 다양한 조사나 접속사로 인해 같은 의미를 구분 짓기 어려운 상황이 많이 존재하여 정확하게 똑같은 표현만을 좋은 표현이라고 단정 짓기 어렵다. 그래서 BLEU를 사용하여 한국어의 번역 품질을 자동평가 할 때 형태소 분석기를 통해 최소 단위로 형태소를 분리해서 평가에 사용한다.

본 연구에서는 특히 문헌의 기계번역 결과에 대한 자동평가에 있어, 단어와 단어사이의 n-gram으로만 평가 하는 BLEU 평가 방식이 아닌, 어순이 평가에 직접적인 영향을 주는 RIBES를 사용하고, 그 기본 평가 방법을 개선하기 위해 복합 형태소 분리를 통한 방식을 제안하고자 한다.

### 2.2. RIBES 한국어 자동평가 방법

RIBES의 평가는 어순과 일치하는 단어의 비교로 이루어진다.

$$NKT = 2 \times \frac{\text{the number of increasing pairs(순위증가 단어의 쌍)}}{\text{the number of all pairs(모든 단어의 쌍)}} - 1$$

여기서 어순의 일치는 모든 단어의 쌍과 그 단어의 쌍 중에서 순서가 증가한 쌍 즉, 순서가 역전되지 않은 정배열 되어있는 쌍의 수로 NKT(Normalized Kendall Tau)를 계산한다. 다음 표 1.의 예시로 설명을 하면,

표 1. NKT 계산 예시 문장[2]

유형	문장
휴번번역 (reference)	John hit Bob yesterday
	[1, 2, 3, 4]
기계번역 (hypothesis)	Bob hit John yesterday
	[3, 2, 1, 4]

평가대상인 hypothesis의 모든 단어의 쌍은 표 2.와 같이 6개 이며, 기계번역 문장의 단어의 등장 순위를 reference문장 순위와 비교해 보면 [3, 2, 1, 4] 순으로 배열되어 있고 단어의 순위증가 쌍은 3개이므로  $NKT = 2 \times 3 / 6 - 1 = 0$ 이다. 즉, 순위의 역전이 많은 문장일수록 평가점수에 미치는 영향이 커진다. 길이가 다른 두 문장일 경우에는 전혀 일치하지 않아 필요 없는 단어를 배제하여 길이를 맞추고 단어 쌍을 찾아 순위변화를 비교한다.

표 2. 단어 쌍의 구분

유형	단어 쌍	
모든 단어 쌍	[Bob, hit], [Bob, John], [Bob, yesterday], [hit, John], [hit, yesterday], [John, yesterday]	
순위 증가 단어 쌍	[Bob, yesterday]	[3, 4]
	[hit, yesterday]	[2, 4]
	[John, yesterday]	[1, 4]

정밀도는 그림 1.과 같이 hypothesis에서 구분된 단어의 개수와 단어의 위치가 완전 일치하거나, 앞 또는 뒤에 동일 단어(bigram window)를 가진 단어의 개수로 계산한다.

$$\text{정밀도} = 1.0 \times \frac{\text{위치, 앞뒤 단어(bigram window)가 동일 단어의 개수}}{\text{hypothesis에서 구분된 단어의 개수}}$$

$$\text{페널티} = \min(1, \exp(1 - \frac{\text{reference에서 구분된 단어의 개수}}{\text{hypothesis에서 구분된 단어의 개수}}))$$

$$RIBES = \frac{NKT+1}{2} \times \text{정밀도}^{0.25} \times \text{페널티}^{0.1}$$

Read a hypothesis sentence  $h = h_1 h_2 \dots h_m$   
and its reference sentence  $r = r_1 r_2 \dots r_n$ .

Initialize worder with an empty list.

For each word  $h_i$  in  $h$ :

- If  $h_i$  appears only once each in  $h$  and  $r$ , append  $j$  s.t.  $r_j = h_i$  to worder.
- Otherwise, if the bigram  $h_i h_{i+1}$  appears only once each in  $h$  and  $r$ , append  $j$  s.t.  $r_j r_{j+1} = h_i h_{i+1}$  to worder.
- Otherwise, if the bigram  $h_{i-1} h_i$  appears only once each in  $h$  and  $r$ , append  $j$  s.t.  $r_{j-1} r_j = h_{i-1} h_i$  to worder.

Return worder.

그림 1. 순위의 상관관계를 위한 단어 정렬 알고리즘[2]

쉽게 말해, 첫 번째 그림 2.와 같이 구분된 단어의 배열을  $h[i] == r[j]$ 로 비교하여 각 하나씩 존재하는 동일 단어일 경우 리스트에 추가한다. [5. | 있어서, | 복수 | 경우,]

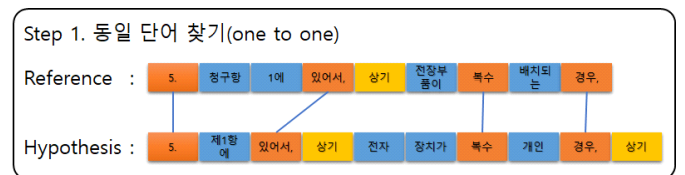


그림 2. 각 문장에 단일 존재하는 동일 단어 찾기

하지만 그림 2.의 “상기”와 같이 두 문장사이에 N개가 존재하면 그림 3.과 같이 정확한 단어 등장순위를 찾

기 위해 앞뒤단어를 비교하여 일치 것이 있으면 동일 단어 리스트에 추가한다. “상기”의 앞뒤단어 비교로 “있어서,”를 가진 앞쪽 “상기”가 동일 단어 리스트에 추가되는 것이다.



그림 3. 각 문장에 N개 존재하는 동일 단어 찾기

일치단어일지라도 앞뒤 단어(bigram window)가 1개라도 일치하지 않는다면, 정밀도계산에서 제외시킨다. 그런 다음 단어의 등장 순서를 비교하여 순위가 증가한 쌍과 모든 단어의 쌍으로 NKT를 계산하여 문장에 RIBES 평가지표를 산출한다.

2.3. RIBES 한국어 자동평가 문제점

RIBES 자동평가에서 띄어쓰기로 분리한 경우 표 3.과 같이 사람이 직접 번역한 휴먼번역(reference)과 기계번역기가 번역한 기계번역(hypothesis)문장의 평가에서 기계번역 1의 경우 수렴 가능한 결과가 나왔지만, 기계번역 2의 경우 의미 없는 평가점수가 나온다.

표 3. 번역문장

유형	문장	평가
휴먼번역 (reference)	5. 청구항 1에 있어서, 상기 전장부품이 복수 배치되는 경우, 직렬 배치되는 것을 특징으로 하는 차량용 냉각회로.	
기계번역 1 (hypothesis)	5. 제1항에 있어서, 상기 전자 장치가 복수 개인 경우, 상기 복수 개의 전자 장치들은 직렬로 배치되는 것을 특징으로 하는 냉각 회로.	76.91
기계번역 2 (hypothesis)	제 1 항에있어서, 상기 전자 장치가 복수로 제공되는 경우, 상기 복수의 전자 장치는 직렬로 배치되는 냉각 회로.	0.0

기계번역 2 문장은 휴먼번역 문장과 비교하여 잘못된 문장이라고 평가하는 것은 무리가 있다. 한국어는 교착어이고, ‘주어의 생략’, ‘서술어의 축소’, ‘외래어 표기방법의 다양성’ 등의 한국어의 특성으로 문법에 맞지 않는 문장이라 할지라도 자연스러운 문장일 수 있다. 특히 특허와 같은 기술문서에서는 잘 사용되지 않는 기술용어들도 존재하고 특허는 등록을 위해 어렵게 풀이 기술하기 때문에 더 많이 발생한다.

3. 제안 방법

3.1. RIBES 한국어 평가 문제 해결방법 제안

RIBES는 기본적으로 띄어쓰기 기반으로 문장을 평가하기 때문에 평가를 위해 한국어에 맞는 형태소 분석기를 사용하여 문장을 새로 분리해 주어야 한다. Workshop on Asian Translation(WAT)에서도 한국어의 자동평가에서

MeCab[5] 형태소 분석기를 사용하여 BLEU, RIBES로 평가하도록 하고 있다[6]. 하지만 기계번역문에서 “항에있어서,”를 보면 한국어 문법에서의 띄어쓰기가 잘못되어 있다. 이를 형태소 분석을 하게 되면 “항”, “에”, “있”, “어서”, “,”로 세분화되어 문법상 잘못된 부분에 대해 잘된 문장과 동일하게 평가하게 된다. Liling Tan et al., 15;에서는 사소한 어휘차이가 불필요할 만큼 많이 분리된 형태소로 인해 n-gram 정밀도에 큰 차이를 일으킨다고 말하고 있다[7].

본 연구에서는 평가문장을 의미와 문법을 해치지 않는 복합 형태소 분석을 통해 형태소를 원래의 띄어쓰기 구조로 유지시켜 특허 기계번역의 한국어 자동평가에 대해 형태소 분석으로 진행한 결과보다 휴먼평가와 유사한 결과를 얻었다.

3.2. 복합 형태소 분석 방법

두 문자간의 비교 평가를 위해 표 1.의 번역문장의 일부분으로 만든 표 4.을 예시로 보면,

표 4. 복합 형태소 분석 예시 문장

유형	문장
휴먼번역 (reference)	5. 청구항 1에 있어서, 상기 전장부품이
기계번역 (hypothesis)	제 1 항에있어서, 상기 전자 장치가

복합 형태소 분석에 앞서 그림 4.와 같이 MeCab-ko[8]를 활용하여 문장을 분리하였다.



그림 4. 형태소 분석(MeCab-ko)

형태소 분석의 경우 그림 4.와 같이 최소단위로 분리되어 일치하는 부분이 상당히 많이 발생한다. 0점이었던 표 3.의 기계번역 2 문장의 RIBES 평가 점수는 81.44로 증가하였다. 불필요할 만큼 많은 형태소로 정밀도를 상승시키는 요소를 배제하기 위해 우리는 형태소 분석에 띄어쓰기 정보를 활용하여 복합 형태소 분석 방법[9]을 사용하여 자동평가를 진행하였다.



그림 5. 복합 형태소 분석

그림 5.과 같이 띄어쓰기 단위로 분리한 다음 그 안에서 의미부와 조사부를 나누어 자동 평가하도록 하였다.

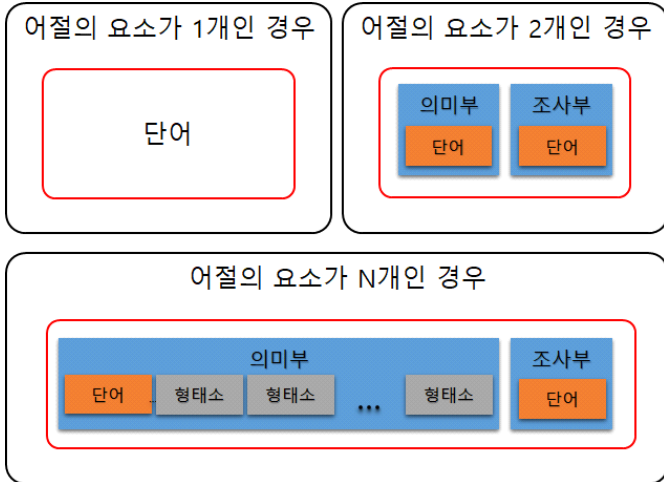


그림 6. 의미부(어두)와 조사부(어미) 구분

복합 형태소는 그림 6.와 같이 문장을 띄어쓰기로 먼저 나눈 어절을 각각의 형태소로 구분하여 의미부와 연결조사부분으로 구분지어 사용한다. 어절의 요소가 1개인 경우 1개의 형태소로만 이루어져 있고, 2개인경우도 단어와 연결조사, 단어와 접미사, 접두사와 단어, 또는 합성어로 이루어진 형태이기 때문에 형태소 분석과 큰 차이는 없다. 하지만 2개 이상인 경우는 띄어쓰기가 잘못되거나 합성어와 조사, 접두사와 단어와 접미사가 붙은 형태로 중간 부분은 불필요한 형태소로 간주한다. 하지만 중간부분 역시 bigram window를 찾을 때는 사용한다.

#### 4. 실험 설계

##### 4.1. 특허 영어문장 데이터

본 실험에서 사용한 기계번역 데이터 셋은 특허의 A-H 섹션 각 10문장, 견해서 20문장으로 구성하여 총 100개의 영어-한국어 번역문장으로 구성하였다. 문장 구성은 길이가 긴/중간/짧은 문장을 같은 비율로 구성하였다.

##### 4.2. 기계번역

번역문장 100개에 대해 구글 번역 API를 이용하여 번역을 진행하였다. 또한 특허 번역 경력이 있는 외부 전문가를 통해 영어원문에 대한 reference 데이터를 구축하였다.

##### 4.3. 휴먼평가

자동평가에 앞서 비교대상군인 사람의 평가지표는 5점 척도로 진행하였다.

표 5. 평가 기준

점수	기준
1	대부분의 정보가 이해 불가능하고 정확히 번역되지 않았다.
2	번역된 정보가 부분적으로 이해가능하게 쓰였다.
3	번역된 대부분의 정보가 이해가능하게 쓰였다.
4	번역된 대부분의 정보가 정확하게 쓰였다.
5	번역된 모든 정보가 정확하게 쓰였다.

평가기준은 표 5.에 맞게 특허 번역 경력이 있는 외부 전문가(4명)를 통해 원문과 기계번역문을 비교 평가하였다.

#### 4.4. 자동평가

자동평가는 NTT의 RIBES 연구진이 공식 배포한[10] Python 프로그램을 활용하였으며, 형태소 분석에는 MeCab을 사용하여 평가를 진행하였다. 복합 형태소 분석에는 띄어쓰기로 분리시킨 어절을 MeCab를 통해 2차 분리를 진행하여 평가 데이터로 사용하였다.

### 5. 연구 결과 및 고찰

실험데이터로 기계 번역한 문장을 휴먼평가와 띄어쓰기, 형태소 분석, 복합 형태소 분석으로 RIBES 자동평가를 진행하여 결과를 특허의 A-H섹션, 견해서 별로 평가 평균으로 그림 7.과 같은 결과를 얻었다.

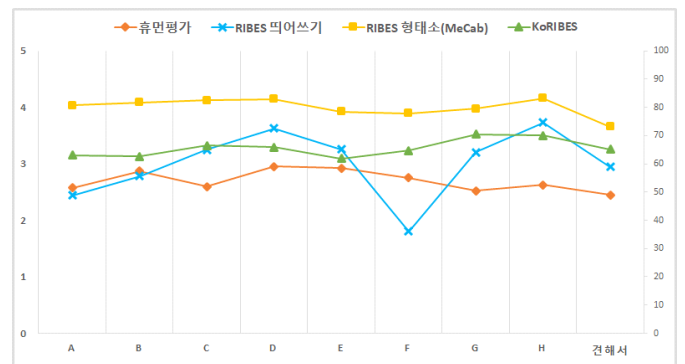


그림 7. 휴먼평가와 RIBES(띄어쓰기/형태소/복합 형태소) 자동평가 결과

그림 7.과 같이 띄어쓰기의 평가결과는 상당히 불규칙적이며, 형태소 분석을 한 경우 휴먼평가에 비해 상당히 높은 평가 점수가 나왔는데 이는 불필요할 만큼 많이 분리된 형태소로 인해 발생한다.

	휴먼평가	p-value
RIBES 띄어쓰기	0.000343972	4.09e-51
RIBES 형태소(MeCab)	<b>0.345767308</b>	2.3e-99
KoRIBES	0.207268215	1.984e-83

그림 8. 휴먼평가에 대한 RIBES 자동평가의 피어슨 상관관계

다음으로 3가지 방법에 대해 그림 8.에서와 같이 피어슨 상관관계를 측정하였고 모두 휴먼평가와의 유의확률에 있어서 p-value는 각 4.09e-51, 2.3e-99, 1.984e-83으로 유의수준 5%이하이기 귀무가설이 기각되어 서로 상관관계가 있다고 할 수 있다.

휴먼평가에 대해 각 방법에 대한 피어슨 상관관계를 보면 MeCab 형태소 분석으로 자동평가를 결과가 제안한 방법보다 높은 양의 상관관계를 가지고 있지만 전체점수가 과도하게 높게 측정되고 문장별로 편차가 제안한 방법보다 크다.

구분	기술 분류								
	A	B	C	D	E	F	G	H	컨해서
RIBES+ 띄어쓰기	-0.29	-0.07	0.22	-0.04	0.26	0.04	-0.22	0.35	0.12
RIBES+ MeCab	0.48	<b>0.69</b>	<b>0.52</b>	-0.28	<b>0.54</b>	-0.04	<b>0.15</b>	0.41	0.28
KoRIBES	<b>0.57</b>	0.67	0.29	-0.26	-0.22	<b>0.32</b>	-0.07	<b>0.42</b>	<b>0.32</b>

그림 9. 휴먼평가에 대한 특허 섹션별 피어슨 상관관계

그리고 그림 9.와 같이 특허에서는 기술 분야별로 신조어, 전문용어, 고유명사의 신규 발생 빈도에 따라 사용되는 용어와 서술방식의 차이가 있고, 이에 따라 기술 분류별 휴먼평가와의 피어슨 상관관계가 기술 분야 마다 차이를 보임을 알 수 있다.

구분	기술 분류									표준 편차
	A	B	C	D	E	F	G	H	컨해서	
휴먼평가	-0.12	0.18	-0.1	0.25	0.23	0.05	-0.17	-0.07	-0.25	0.173
RIBES+ 띄어쓰기	-0.56	-0.22	0.25	0.62	0.26	-1.2	0.2	0.73	-0.06	0.566
RIBES+ MeCab	0.03	0.08	0.13	0.15	-0.07	-0.11	-0.02	0.16	-0.35	0.152
KoRIBES	-0.13	-0.15	0.05	0.02	-0.19	-0.04	0.24	0.22	-0.02	<b>0.144</b>

그림 10. 자동평가 결과의 표준편차(STDEVP)

또한, 자동평가 결과를 그림 10.과 같이 표준편차(STDEVP)로 비교했을 때 제안한 복합 형태소 분석으로 평가한 결과가 띄어쓰기로 분리한 것과 형태소 분석으로 분리하여 평가한 것보다 상대적으로 가장 적은 편차의 결과를 보임을 알 수 있다.

## 6. 결론 및 향후 방향

본 연구는 기계번역의 자동평가에서 기계번역이 많이 사용되는 특허 번역에 대해 영어-한국어 번역 결과물의 평가에 효율적인 방법을 제시하고자 진행하였다. 연구결과에서와 같이 제안한 방법이 표준편차에서는 안정적인 결과를 보였고, 피어슨 상관관계에서는 형태소 분석(MeCab)이 좋은 결과를 보였으나, 기술 분류에 따라 다른 결과를 보이거나, 평가 결과의 수준이 휴먼평가에 비해 과도하게 평가되어 형태소 분석만을 사용한 평가 방식이 복합 형태소 분석에 의한 평가 방식이 RIBES를 활용한 평가에 있어 보다 적절하다고 판단된다.

특허 기계번역에서 영어→한국어로 기계 번역된 결과물의 RIBES 자동평가에서 tokenization에 따른 평가척도의 정확성개선을 위해 휴먼평가의 데이터가 신뢰할 수준의 많은 데이터를 수집하고 특허 문서에서 한국어가 가지는 특성과 신규단어의 발생 등을 고려하여 대규모 말뭉치를 BPE(Byte Pair Encoding; Sennrich et al., 16)[11]로 sub-word tokenize하는 등의 여러 실험을 통해 자동평가 정확성 향상을 위한 후속 연구가 필요하다.

기계번역의 오역과 불완전성을 없애기 위해 번역 품질에 대한 평가는 필수적인 요소이며, BLEU, sacreBLEU, RIBES, METEOR, AMFM 등과 같은 좋은 평가방법이 연구되

고 언어에 맞는 평가 방법을 적절히 잘 활용한다면 기계번역의 품질은 빠르게 발전할 것이고, 기계번역을 통해 전 세계 특허나 논문 같은 기술문서를 해당 언어에 대한 지식 없이 잘 이해 할 수 있는 날이 오기를 기대해본다.

## 참고문헌

- [1] 최효은, 이지은, “특허 기계번역 결과물의 평가 - KIPRIS의 무료 한영 기계번역을 중심으로”, 2017
- [2] Kishore Papineni, Salim Roukos, Todd Wart, and Wei-Jing Zhu, “BLEU : a Method for Automatic Evaluation of Machine Translation”, 2002.
- [3] 특허청, 한국지식재산서비스협회, “지식재산 번역 가이드라인”, 2010.
- [4] Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada, “RIBES : Automatic Evaluation of Translation Quality for Distant Language Pairs”, 2010.
- [5] Taku Kudou, “MeCab : Yet Another Part-of-Speech and Morphological Analyzer”, 2005.
- [6] Chenchen Ding, Masao Utiyama, Eiichiro Sumita, “Structured Common Subsequences for Automatic Machine Translation Evaluation”, 2018.
- [7] Liling Tan, Jon Dehdari, Josef van Genabith, “An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation”, 2015.
- [8] 은전한빛 MeCab-ko, <http://bitbucket.org/eunjeon/mecab-ko/src/master>
- [9] 양성일, 홍문표, 김영길, 최승권, “띄어쓰기 정보를 이용한 한국어 복합 형태소 분석”, 2003.
- [10] NTT RIBES Open Source, <http://www.kecl.ntt.co.jp/icl/lirg/ribes>
- [11] Rico Sennrich, Barry Haddow, Alexandra Birch, “Neural Machine Translation of Rare Words with Subword Units”, 2016.