http://dx.doi.org/10.15749/jts.2020.21.5.012

번역학연구
2020 겨울호 제21권5호

NMT의 평가항목별 자동평가를 위한 영한 평가세트 연구*

최승권[1] · 한지은[2] · 최규현[1] · 김영길[1]
(한국전자통신연구원[1] · 한국외국어대학교[2])

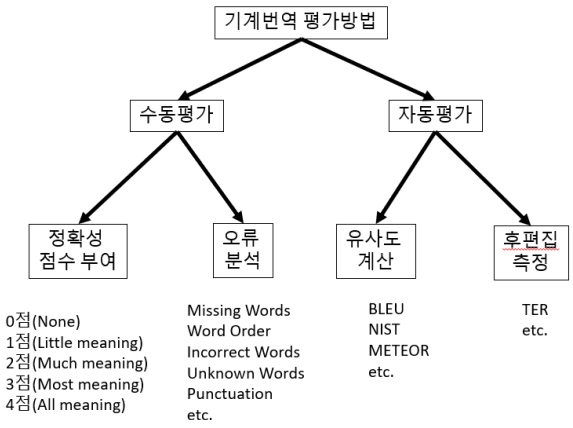
1. 서론

2016년 11월에 Google이 기계번역에 적용되는 기술을 통계 기반에서 인공 신경망 기반으로 전환하였다(Wu et al. 2016). 그 결과 인공 신경망 기반의 신경망 기계번역(이후로 NMT로 표현하겠다)의 성능은 통계 기반의 기계번역 성능을 앞지르고 있다고 보고되었다(Koehn and Knowles 2017). 대용량의 번역 말뭉치가 누적되어 심층 학습되면 될수록, NMT의 번역 품질은 계속 향상되고 있다. 하지만 NMT의 번역품질이 아무리 개선되고 있어도 인간 번역사가 보기에는 여전히 부족한 점이 많다. NMT가 시간과 비용의 경제성, 접근의 용이성과 사용의 편리성은 있지만, 문맥 파악이나 생략 표현이나 노이즈 표현 등으로 인해 전혀 이해할 수 없는 오역이 나오는 단점이 있다고 평가하였다(강수정 2020).

* 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

기계번역 시스템을 평가하는 방법은 기계번역 결과를 인간 번역사가 평가하는 수동평가 방법과 BLEU(Papineni et al. 2002)와 같은 자동평가 방법으로 나눌 수 있겠다.

그림 1 기계번역 평가 방법의 분류



수동평가 방법은 평가의 기준이 무엇인가에 따라 정확성 점수(accuracy score) 부여 방법, 오류 분석(error analysis) 방법 등으로 나눌 수 있다. 정확성 점수 부여 방법은 자동 번역된 결과에 번역사가 평가 점수를 부여하는 것이다. 정확성 점수 부여 방법은 원문의 의미를 번역문에 어느 정도로 전달하였는지를 평가하는 방법으로 대개 0~4점을 부여하고 평가자의 최고/최저 점수를 뺀 평균 점수로 평가를 하는 방법이다. 사용되는 점수는 0점(None), 1점(Little meaning), 2점(Much meaning), 3점(Most meaning), 4점(All meaning)으로 나뉜다(Doyon et al 1998). 오류 분석 방법은 기계번역 결과를 인간 번역사가 분석하여 오류의 유형을 분류함으로써 기계번역 시스템의 개선 방향을 제시하려는 것이 목적이다. 오류 분석 방법은 본 논문의 주제와도 관련이 있으므로 2장에서 더 자세하게 소개하도록 하겠다.

자동평가 방법은 기계번역 결과와 인간 번역사의 정답문(reference)과의 유사도(similarity)를 계산하는 방법으로 BLEU, NIST, METEOR 등이 있으며, 기계번역 결과의 후편집(post-editing)과 정답문과의 비교를 자동으로 측정하는

TER(Snover et al. 2006)방법으로 나눌 수 있겠다.

수동평가 방법은 정확성은 있으나 인간 번역사에 의한 평가이므로 시간과 비용이 많이 든다는 단점이 있다. 자동평가 방법은 정답문만 존재하면 NMT 시스템을 자동으로 평가할 수 있다는 장점은 있으나, 평가 점수만 보고서는 수동평가 방법의 오류 분석과 같은 상세한 분석을 알 수 없다는 단점이 있다. 따라서 기존의 자동평가 방법으로는 NMT 시스템의 어느 부분이 장점이고 어느 부분이 단점인지를 파악하기 어렵다.

이러한 NMT 시스템의 평가 문제점을 개선하기 위해 본 논문에서는 NMT의 장단점을 자동으로 평가하며 수동평가 방법의 오류 분석처럼 NMT의 오류를 직관적으로 파악할 수 있는 평가항목별 자동평가 방법을 제안하고자 한다. 평가항목별 자동평가 방법을 제안하기 위해 NMT를 위한 평가항목별 분류체계와 분류체계에 따른 평가항목별 자동 평가 방법을 기술할 것이다. 이를 위해 본 논문은 다음과 같이 구성된다. 2장에서는 NMT의 오류 분석에 대한 선행연구를 기술하겠다. 3장에서는 자동 평가용 평가 세트를 기술한다. 4장에서는 NMT 시스템의 평가항목별 자동평가 방법에 대해 기술한다. 5장에서는 국내에서 잘 알려진 2개의 NMT 영한 시스템에 대한 평가항목별 자동평가 결과를 밝히고자 한다.

2. 기존의 기계번역 오류 분석 방법

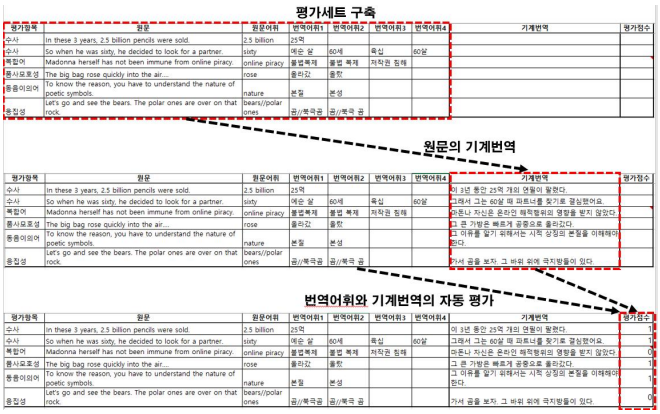
기계번역 시스템에 대한 오류 분석 방법은 번역 시스템의 개발 역사와 더불어 다양하게 발전하였다. 기계번역 시스템들의 다양한 오류 유형을 제시하려는 오류 분석 연구(Villar et al. 2006; Popović and Ney 2011)로부터, 오류 유형을 이용하여 기계번역 시스템들의 성능을 비교하려는 연구(Bojar 2011; Costa et al 2015; Bentivogli et al 2016; Isabelle et al 2017; Machetanz et al 2018), 오류 유형을 부착한 말뭉치를 구축하여 기계번역 성능을 자동으로 측정하려는 연구(Fishel et al 2012; Wisniewski et al 2014), 기계번역과 후편집의 오류 유형을 비교하는 연구(Daems et al 2014), 기계번역 결과를 정확성(accuracy)과 유창성(fluency)으로 구분하여 오류 유형을 설명하려는 연구(Lommel et al. 2014;

서보현·김순영2018; Brussel et al. 2018), BLEU와 같은 자동 평가 방법을 인간 번역 평가에 활용하려는 연구(정혜연 2018; 김보영 2020) 등이 있었다. 또한 특정 오류 유형을 지정하여 상세하게 분석한 연구가 있었는데, 대명사 번역 오류 분석 연구(Guillou & Hardmeier 2016), 무생물 주어 번역 오류 분석 연구(최동익 2013) 등이 있었다.

3. 자동평가용 영한 평가세트와 자동평가 방법

본 장에서는 NMT 시스템을 평가항목별로 자동으로 평가할 수 있는 평가항목별 자동 평가용 평가 세트[1]에 대해 기술하고자 한다. 여기서 평가 세트란 NMT 시스템을 평가하기 위한 데이터 집합을 말하는 것으로 평가항목, 원문, 원문어휘, 번역어휘, 기계번역, 평가점수가 평가를 위한 하나의 세트(set)로 구성되어 있어 평가 세트라는 표현을 사용하고자 한다. 아래의 그림과 같이 1) 평가세트 구축, 2) 원문의 기계번역, 3) 번역어휘와 기계번역의 자동 평가 순으로 설명하도록 하겠다.

그림 2 자동평가용 영한 평가세트 구축과 자동평가 절차



1) 평가 세트는 영미권에서는 test suite라는 용어로 사용되곤 한다.

3.1 평가 세트 구축

3.1.1 평가 세트의 구조

평가 세트를 구성하는 요소들은 다음과 같다.

〈표 1〉 자동 평가용 평가 세트의 구조

평가항목	원문	원문어휘	번역어휘	기계번역	평가점수

평가항목 : 자동 평가를 위한 항목의 이름. (자세한 내용은 3.1.2 절에서 상세히 기술하겠다)

원문 : 평가 대상이 되는 출발 언어의 문장 또는 문장들.

원문어휘 : 원문에서 평가하고자 하는 평가항목의 어휘.

번역어휘[+2)]: 원문어휘에 등가되는 번역 어휘.

기계번역 : NMT 시스템이 원문을 번역한 결과.

평가점수 : NMT 결과에 대해 번역어휘를 가지고 자동평가한 점수.

각각의 평가항목은 올바르게 평가할 수 있기 위해 원문, 원문어휘, 번역어휘, 기계번역, 평가점수를 10개 이상씩 가지도록 하였다. 원문은 평가항목에 따라 한 문장이 될 수도 있고 여러 문장이 될 수도 있다. 여러 문장인 경우는 문장간 응집성을 평가하는 경우가 이에 해당한다. 원문어휘는 한 단어이거나 여러 단어일 수 있다. 번역어휘는 원문어휘에 등가되는 번역을 말하며 원문의 문맥을 반영한 등가어가 되어야 하므로 자연스럽게 문맥도 반영하는 효과를 가진다. 번역어휘는 그 형태와 수에 따라 다음과 같이 다양하게 기술할 수 있다:

1) 번역어휘가 1개인 경우

설명) 평가 세트의 기본 구조로, 원문어휘에 1개의 번역어휘가 등가되는 경우이다.

2) <번역어휘>의 오른쪽에 붙은 +는 <번역어휘>가 1개 이상이어야 한다는 것을 의미한다.

예)

평가 항목	원문	원문 어휘	번역 어휘
수사	In these 3 years, 2.5 billion pencils were sold	2.5 billion	25억

2) 번역어휘가 여러 개인 경우

설명) 원문어휘에 번역어휘가 2개 이상 대응되는 경우이며 콤마로 구분한다.

예)

평가 항목	원문	원문 어휘	번역 어휘
동음이의어	To know the reason, you have to understand the nature of poetic symbols.	nature	본질, 본성

3) 번역어휘가 NMT 결과인 경우

설명) NMT 결과 중에 번역어휘에 추가할 수 있는 경우, 번역어휘 앞에 # 를 표시하여 구별한다.

예)

평가 항목	원문	원문 어휘	번역 어휘
숙어	Whether or not the economy is entering a recession is really beside the point.	b e s i d e the point	주제와 관계없, #요점을 벗어

4) 번역어휘가 분리될 수 있는 경우

설명) 원문어휘에 등가되는 번역어휘가 분리되어 나타날 수 있을 경우, 분리되는 부분에 //를 표시하여 구별한다.

예)

평가 항목	원문	원문 어휘	번역 어휘
대명사	None of the books were interesting	none	어느//지 않, 어떤//지 않

5) 번역어휘가 되지 말아야 하는 경우

설명) 번역어휘가 번역문에 나타나지 말아야 하는 경우, 번역어휘 앞에 ~를 표시하여 구별한다. '~'은 논리적 부정(negation)을 의미한다.

예)

평가 항목	원문	원문 어휘	번역 어휘
대명사	The party went to his rescue.	his	그를, ~그의

3.1.2 평가 세트의 평가 항목

NMT 시스템의 장단점을 파악하기 위하여 다양한 평가항목들을 수집하고 선정하려고 하였다. 그래서 2장에서 소개된 기존의 기계번역 오류 분석의 오류 유형들을 평가항목들로 수집하였다. 중복된 오류 유형을 포함하여 오류 유형은 총 222개로 파악되었다. 이 222개의 오류 유형은 대부분 영어 관련 오류 유형이었지만 '무생물 주어'와 같은 한국어 관련 오류 유형들도 포함하고 있었다. 이 222개로부터 최종적인 평가항목들을 선정하기 위해 다음과 같은 오류 유형들을 제거하였다

- 'Multiword Expression' 또는 'Multi word syntax' 등과 같은 중복 표현.
- 'content word' 또는 'phrase level'과 같은 넓은 범주 표현.
- 어순(word order) 오류나 번역되지 않은(untranslated / not translated) 오류와 같은 번역문에만 나타나는 오류 유형.

번역문에만 나타나는 오류 유형이 제거된 이유는 원문만으로 이런 오류 유형의 발생을 예측할 수 없었기 때문이다. 이런 제거 과정을 통하여 47개의 최종적인 평가항목을 얻었다. 평가항목을 일목요연하게 파악하기 위해 언어학적 문법 체계에 따라 분류하였다. 문법 체계는 형태론(morphology), 구문론(syntax), 의미론(semantics), 담화(discourse)로 범주(category)를 설정하고 다시금 하위 범주(subcategory)로 세분화하였다. 이와 같은 체계 하에 평가항목은 다음과 같은 분류 체계를 가지게 되었다.

〈표 2〉 자동 평가용 평가항목 분류 체계

범주	하위범주	평가항목
형태론	품사	관사, 명사, 대명사, 형용사, 부사, 전치사, 동사, 조동사, 관계대명사, 관계부사, 접속사, 기호, 수사
구문론	단어	복합어
	구	부정사구문, 분사구문, 동명사, 비교구문, 가정법, 수동구문, 삽입, 생략/도치, 병렬, 수량표현, 음성문자
	결합어	숙어, 동사+전치사, 연어
	문장 유형	의문문, 명령/감탄/청유문, 부정문, 무생물주어문
	장문	20<=단어수<30, 30<=단어수<40, 40<=단어수
	모호성	품사모호성, 구조모호성
의미론	동음이의어	동음이의어
	화법	화법, 스타일
	시제/상	시제/상
	개체명	인명, 도시, 조직, 시간, 장소
담화	응집성	응집성
4개	12개	47개

3.1.3 평가 세트의 구축 방법

평가항목에 대한 분류 체계가 완료된 후, 평가항목에 맞는 영어 원문과 영어 원문어휘, 한국어 번역어휘를 구축하였다. 원문의 수집, 원문어휘의 선정, 번역어휘의 구축과 검증 방식은 다음과 같았다:

- 말뭉치: 한국전자통신연구원(ETRI)에 구축되어 있는 약 1,000만 문장 이상의 영한 번역말뭉치를 대상으로 하였다.
- 원문 수집: 한국전자통신연구원(ETRI)에서 인턴으로 근무한 영어 전공 4학년 학생 2명이 평가항목에 대한 교육을 받은 후 1개월에 걸쳐 영어 원문을 수집하였다.
- 원문어휘 선정 및 번역어휘 구축: 영어 원문을 수집한 영어 전공 4학년 학생 2명이 영어 원문에 포함된 평가 항목의 원문 어휘를 선정하고 원문 어휘에 대응되는 한국어 번역 어휘를 구축하였다. 한국어 번역 어휘는 영한 번역말뭉치의 한국어 번역문으로부터 1개를 선정하고 추

가 가능한 어휘는 인터넷 사전으로부터 원문의 문맥을 고려하여 번역 어휘로 선택하였다.

- 검증 방식: 검증은 2단계에 걸쳐 수행하였다. 1단계 검증은 원문 및 원문 어휘가 구축되고 6개월 후에 한국전자통신연구원(ETRI)에 인턴으로 근무하게 된 새로운 영어 전공 학생 2명(기존 구축자와 다른 학생들)이 평가항목에 대한 교육을 받은 후 기존에 구축된 원문 및 원문어휘, 번역어휘를 대상으로 1개월에 걸쳐 검증하였다. 2단계 검증은 1차 검증 결과를 대상으로 평가항목과 원문어휘의 관계성, 원문어휘가 원문에 포함되어 있는지의 여부 등을 본 저자와 1단계 작업을 한 학생들이 0.5개월에 걸쳐 검증하였다.

전체적으로 평가 세트는 2.5개월에 걸쳐 구축되었다.

3.1.4 평가 세트의 통계

NMT 평가를 위한 평가 세트의 평가항목 수는 총 47개였으며 문장은 910문장이 수집되었다. 평가항목별로 수집한 문장은 다음과 같았다:

〈표 3〉 평가항목의 항목수 및 문장수

범주	하위범주	평가항목	문장수
형태론	품사	13	280
구문론	단어	1	20
	구	11	250
	결합어	3	60
	문장 유형	4	60
	장문	3	60
	모호성	2	40
의미론	동음이의어	1	20
	화법	2	30
	시제/상	1	30
	개체명	5	50
담화	응집성	1	10
계		47	910

담화의 응집성 항목의 원문은 여러 문장으로 구성되었지만 1개의 문장으로 간주하여 통계를 내었다.

3.2 원문의 기계 번역

NMT 시스템을 평가하기 위해, 원문의 기계번역은 국내에서 잘 알려진 Naver의 Papago와 Google의 Google Translate NMT 시스템으로 만들었다. (P: Papago, G: Google Translate)

예)

평가 항목	원문	원문 어휘	번역 어휘	기계 번역
수사	In these 3 years, 2.5 billion pencils were sold	2.5 billion	25억	P: 이 3년 동안 25억 개의 연필이 팔렸다.
				G: 3 년 동안 25 억 개의 연필이 팔렸습니다

3.3 번역어휘와 기계번역의 자동 평가

번역어휘와 기계번역 결과를 토대로 자동 평가의 전체적인 절차는 다음과 같다.

〈표 4〉 평가항목별 자동평가 절차

1) 평가 세트의 <번역어휘>와 <기계번역>을 비교한다.
   가) <번역어휘>가 <기계번역>에 존재하면 1점을 부여한다.
   나) <#번역어휘>가 <기계번역>에 존재하면 0.5점을 부여한다.
   다) <번역어휘1//번역어휘2>가 <기계번역>에 존재하면 1점을 부여한다.
   라) <~번역어휘>가 <기계번역>에 존재하면 -1점을 부여한다.
   마) 그렇지 않다면, 0점을 부여한다.
2) 부여된 점수를 가지고 평가항목별, 하위범주별, 범주별 정확률을 구한다.

위의 평가항목별 자동평가 절차에 따라 3.1.1 절에서 소개된 예문들의 기계번역에 평가점수를 자동으로 부여하면 다음과 같다. (P: Papago, G: Google

Translate)

가) <번역어휘>가 <기계번역>에 존재하여 1점을 부여한 경우.

평가 항목	원문	원문 어휘	번역 어휘	기계 번역	평가 점수
동음이의어	To know the reason, you have to understand the nature of poetic symbols.	nature	본질, 본성	P: 그 이유를 알기 위해서는 시적 상징의 본질을 이해해야 한다	1
				G: 이유를 알기 위해서는 시적 상징의 본질을 이해해야합니다.	1

나) <번역어휘>가 <기계번역>에 존재하지 않아 0점을 부여한 경우.

평가 항목	원문	원문 어휘	번역 어휘	기계 번역	평가 점수
숙어	Whether or not the economy is entering a recession is really beside the point.	beside the point	주제와 관계없, #요점을 벗어	P: 경제가 침체기에 접어들고 있는지 아닌지는 정말 요령부득이다.	0
				G: 경제가 불황에 접어 들고 있는지 여부는 정말 중요하지 않습니다.	0

다) <~번역어휘>가 <기계번역>에 존재하여 -1점을 부여한 경우.

평가 항목	원문	원문 어휘	번역 어휘	기계 번역	평가 점수
대명사	The party went to his rescue.	his	그를, ~ 그의	P: 일행은 그를 구하러 갔다.	1
				G: 파티는 그의 구조에 갔다.	-1

평가 점수에 따라 평가항목별, 하위범주별, 범주별로 정확률을 구하는 통계는 다음과 같은 식에 의해 구하였다:

$$\text{평가항목 정확률} = \frac{\text{평가항목 문장의 자동평가점수의 합}}{\text{평가항목의 문장수}}$$

$$\text{하위범주 정확률} = \frac{\text{하위범주별 평가항목 정확률의 합}}{\text{하위범주의 평가항목 수}}$$

$$\text{범주 정확률} = \frac{\text{범주별 하위범주 정확률의 합}}{\text{범주의 하위범주 수}}$$

4. 평가 실험

실험에 사용된 NMT 시스템은 앞서 언급한 것처럼 국내에서 잘 알려진 Naver의 Papago와 Google의 Google Translate 시스템이었으며 언어쌍은 영한을 대상으로 평가하였다. NMT 시스템들의 평가 결과는 다음과 같았다. (P: Papago, G: Google Translate)

〈표 5〉 NMT 시스템의 평가항목별 정확률

범주			하위범주			평가항목		
	P	G		P	G		P	G
형태론	83.21	73.75	품사	83.21	73.75	관사	80.00	80.00
						명사	77.50	80.00
						대명사	80.00	78.33
						형용사	85.00	68.33
						부사	85.00	87.50
						전치사	82.50	75.00
						동사	85.83	75.00
						조동사	90.00	70.00
						관계대명사	75.00	35.00
						관계부사	90.00	90.00
						접속사	85.00	87.50
						기호	65.00	70.00
						수사	86.67	56.67
구문론	77.86	62.14	단어	85.00	75.00	복합어	85.00	75.00
			구	74.60	53.80	부정사구문	83.75	67.50

						분사구문	81.67	50.00
						동명사	65.00	50.00
						비교 구문	85.00	65.00
						가정법	95.00	75.00
						수동구문	80.00	60.00
						삽입	75.00	55.00
						생략/도치	80.00	70.00
						병렬	90.00	75.00
						수량표현	80.00	75.00
						음성문자	63.50	40.00
			결합어	84.17	77.50	동사+전치사	93.33	81.67
						숙어	65.00	70.00
						연어	80.00	75.00
			문장 유형	76.67	68.33	의문문	80.00	65.00
						명령/감탄/청유문	70.00	65.00
						부정문	95.00	95.00
						무생물주어문	71.67	61.67
			장문	82.50	76.67	20<=단어수<30	82.50	87.50
						30<=단어수<40	80.00	75.00
						40<=단어수	85.00	67.50
			모호성	80.00	53.75	품사 모호성	87.50	67.50
						구조 모호성	72.50	40.00
의미론	91.15	73.08	동음이의어	92.50	90.00	동음이의어	92.50	90.00
			화법	80.00	68.33	화법	100.00	100.00
						스타일	70.00	52.50
			시제/상	88.33	53.33	시제/상	88.33	53.33
			개체명	99.00	81.00	인면	100.00	100.00
						도시	95.00	85.00
						조직	100.00	85.00
						시간	100.00	40.00
						장소	100.00	95.00
담화	40.00	40.00	응집성	40.00	40.00	응집성	40.00	40.00
계							73.06	62.24

표 5의 평가항목별 정확률에 따르면 Papago와 Google의 영한 평가항목별 전체 정확률은 73.06% 대 62.24%로 Papago의 정확률이 10.82% 높았다. 평가항목별 정확률은 원문의 의미를 번역문에 어느 정도로 전달하였는지를 측정하는 기계번역의 수동평가 정확률(accuracy) 측정 방법과 매우 유사하다. 0~4점을

부여하는 수동평가 정확률에서는 4점 만점에 3점 이하(3/4 = 75%)를 의미가 충분히 전달되지 못하는 수준으로 간주한다. 따라서 평가항목별 정확률에서도 75% 이하를 받은 평가항목들을 해당 NMT 시스템의 약점이라고 간주할 수 있다. 표 5에 따르면, Papago의 영한 NMT는 동명사(65.00%), 음성문자(63.50%), 명령/감탄/청유문(70.00%), 무생물주어문(71.67%), 구조 모호성(72.50%), 스타일(70.00%), 응집성(40.00%)들이 약점으로 파악되었다. 이에 반해, Google 영한 NMT는 형용사(68.33%), 조동사(70.00%), 관계대명사(35.00%), 기호(70.00%), 수사(56.67%), 부정사구문(67.50%), 분사구문(50.00%), 동명사(50.00%), 비교구문(65.00%), 수동 구문(60.00%), 삽입(55.00%), 생략/도치(70.00%), 음성문자(40.00%), 숙어(70.00%), 의문문(65.00%), 명령/감탄/청유문(65.00%), 무생물주어문(61.67%), 40단어 이상 장문(67.50%), 품사 모호성(67.50%), 구조 모호성(40.00%), 스타일(52.50%), 시제/상(53.33%), 시간(40.00%), 응집성(40.00%)에 약점을 가지고 있는 것으로 파악되었다.

5. 결론

본 논문에서는 신경망 기계번역(NMT) 시스템을 평가항목별로 자동으로 평가하는 방법을 소개하였다. 기존의 자동평가 방법들이 NMT 시스템의 평가항목별 장단점을 파악하지 못하는 데 반해, 본 논문의 평가항목별 자동 평가 방법에 의하면 NMT 시스템의 평가항목별 장단점이 직관적으로 파악될 수 있었다. NMT 시스템의 평가항목별 자동평가 방법은 평가항목, 원문, 원문어휘, 번역어휘, 기계번역, 평가점수로 구성되는 평가 세트를 토대로, 번역어휘가 기계번역에 존재하는 지의 여부에 따라 자동으로 평가점수가 계산되었다. 자동 평가 방법은 다음과 같았다: 1) <번역어휘>가 기계번역에 있으면 1점을 부여한다. 2) <#번역어휘>가 기계번역에 있으면 0.5점을 부여한다. 3) <번역어휘1//번역어휘2>가 기계번역에 있으면 1점을 부여한다. 4) <~번역어휘>가 기계번역에 있으면 -1점을 부여한다. 5) 그렇지 않다면, 0점을 부여한다.

평가항목별 자동평가 방법을 Naver의 Papago와 Google의 Google Translate NMT 시스템에 적용하여 각 시스템의 장단점을 파악하여 보았다. Papago 영한

NMT의 최대 약점은 응집성(40.00%)의 번역이었으며, Google 영한 NMT의 최대 약점은 관계대명사(35.00%), 음성문자(40.00%), 구조 모호성(40.00%), 시간(40.00%), 응집성(40.00%)의 번역이라는 것을 알 수 있었다.

평가항목별 자동평가 방법의 궁극적인 목적은 NMT 시스템의 다양한 약점을 찾아내어 약점과 관련된 코퍼스를 반자동으로 구축하고 학습하여 NMT 시스템의 번역 성능을 점증적으로 향상시키는 것이다. 본 논문이 NMT 시스템의 장단점을 평가항목별로 자동으로 파악할 수 있다는 장점은 있지만 그럼에도 불구하고 본 논문이 가지고 있는 번역어휘와 기계번역 결과의 일치(matching) 측정이라는 단순화된 자동평가 접근법은 개선되어야 할 여지가 있다. 그런 점에서, 본 논문이 향후에 개선하고자 하는 방향은 1) 영한 이외의 다른 언어쌍으로 평가항목을 확대하기, 2) 평가항목에 따라 원문을 반자동으로 수집하기, 3) 음성을 대상으로 한 자동통역으로 확대하기, 4) 평가항목에 인간번역가가 고려하는 평가항목 포함하기 등이다.

참고문헌

강수정 (2020)「전문번역사들의 NMT에 대한 인식과 수용에 대한 연구 －심층 인터뷰를 중심으로」,『번역학연구』21(3): 9-35.

김보영, 김연주, 서승희, 송신애, 이진현, 전경아, 최지수, 홍승빈, 정혜연 (2020)「번역자동평가에서 풀리지 않은 과제」,『번역학연구』21(1): 9-29.

서보현, 김순영 (2018)「기계번역 결과물의 오류유형 고찰」,『번역학연구』19(1): 99-117.

정혜연 (2018)「번역의 자동평가: 기계번역 평가를 인간번역 평가에 적용해 보기」,『통번역학연구』22(4): 265-288.

최동익 (2013)「무생물 주어 구문에 대한 영한 기계 번역 오류분석」,『언어학연구』29: 279-299.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016) 'Neural versus phrase-based machine translation quality: a case study', *In Proceedings of the 2016 Conference on Empirical Methods in Natural*

Language Processing. Association for Computational Linguistics, Austin, Texas: 257 – 267.

Bojar O (2011) 'Analysing Error Types in English-Czech Machine Translation', *The Prague Bulletin of Mathematical Linguistics,* 63-76.

Brussel, Laura Van, Arda Tezcan and Lieve Macken (2018) 'A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch', *In Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. 3799-3804.

Costa, Angela, Wang Ling, Tiago Luíıs, Rui Correia, Luísa Coheur (2015) 'A Linguistically Motivated Taxonomy for Machine Translation Error Analysis', *Machine Translation* 29(2): 127-161.

Daems, Joke, Lieve Macken and Sonia Vandepitte (2014) 'On the origin of errors: a fine-grained analysis of MT and PE errors and their relationship', *In Proceedings of the Ninth International Conference on Language Resources and Evaluation*. 62-66.

Doyon, J., Taylor, K., and White, J.S. (1998) 'The DARPA MT evaluation methodology: Past and present', *Proceedings of the Association for Machine Translation in the Americas* 1988, 1-4.

Fishel, Mark, Ondřej Bojar, Maja Popović (2012) 'Terra: a Collection of Translation Error-Annotated Corpora', *In Proceedings of the Eighth International Conference on Language Resources and Evaluation*. 7-14.

Guillou, L. and Hardmeier, C. (2016) 'PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation', *In Proceedings of the Tenth International Conference on Language Resources and Evaluation*: 636-643.

Isabelle, P., Cherry, C., and Foster, G. (2017) 'A Challenge Set Approach to Evaluating Machine Translation', *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2486-2496.

Koehn, Philipp, Rebecca Knowles (2017) 'Six Challenges for Neural Machine Translation', *Proceedings of the First Workshop on Neural Machine*

Translation, 28-39.

Lommel, A, A Burchardt, M Popovic, K Harris, E Avramidis, H Uszkoreit (2014) Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data. *EAMT-2014*, 165-172.

Macketanz, Vivien, Eleftherios Avramidis, Aljoscha Burchardt, Hans Uszkoreit (2018) 'Fine-grained evaluation of German-English Machine Translation based on a Test Suite', *Proceedings of the Third Conference on Machine Translation (WMT)*, 2 (Shared Task Papers): 578-587.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002) 'BLEU: a method for automatic evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics', *Association for Computational Linguistics, Stroudsburg,* PA, USA, ACL '02, 311–318.

Popović, Maja and Hermann Ney (2011) 'Towards automatic error analysis of machine translation output', *Computational Linguistics*, 37(4): 657-688.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul (2006) 'A Study of Translation Edit Rate with Targeted Human Annotation', *In Proceedings of Association for Machine Translation in the Americas*. 223-231.

Vilar D, Xu J, D'Haro LF, Ney H (2006) 'Error Analysis of Machine Translation Output', *In: International Conference on Language Resources and Evaluation*, Genoa, Italy, 697-702.

Wisniewski, Guillaume, Natalie Kübler, François Yvon (2014) 'A Corpus of Machine Translation Errors Extracted from Translation Students Exercises', I*n Proceedings of the Ninth International Conference on Language Resources and Evaluation*. 3585-3588.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto

Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2016) 'Google's neural machine translation system: Bridging the gap between human and machine translation.' *CoRR abs/1609.08144. http://arxiv.org/abs/1609.08144.pdf. 39*

[Abstract]

A Study on English-to-Korean Test Suites for NMT Automatic Evaluation by Linguistic Assessment Items

Sung-Kwon Choi[1], Ji-Eun Han[2], Gyu-Hyeun Choi[1] and Youngkil Kim[1]
(Electronics and Telecommunications Research Institute[1] · Hankuk University of Foreign Studies[2])

This paper describes an approach to automatically evaluate Neural Machine Translation(NMT) systems by linguistic assessment items. While the previous automatic evaluation approaches cannot identify the strengths and weaknesses of NMT systems for each linguistic assessment item, our automatic evaluation approach can intuitively determine both strengths and weaknesses of each linguistic assessment item. The automatic evaluation by linguistic assessment items of NMT systems is evaluated based on whether the answer of translation exists in the machine translation results, after building the test suites of the source text, the expressions in the source text, and the translated word.

As applying the automatic evaluation approach by linguistic assessment items to NMT systems of Papago by Naver and Google Translate by Google, we figured out the strengths and weaknesses of each system. The biggest weakness of Papago English-to-Korean machine translation system is Cohesion(40.00%). The most serious weak points of Google English-to-Korean translation system are the translation of Relative pronoun(35.00%), Spoken expression(40.00%), Structural Ambiguity (40.00%), and Cohesion(40.00%).

The main purpose of automatic evaluation by the linguistic assessment items is to find various weaknesses of the machine translation systems, semi-automatically collect and build a targeted corpus based on the weaknesses, and improve the performance of the machine translation systems incrementally by retraining. Although this paper has the advantage of automatically

recognizing the strengths and weaknesses by linguistic assessment items, the simplified automatic evaluation approach, a measurement based on the matching of translated word and machine translation, that this paper suggests should be improved. In this respect, the improvement directions of this paper in the future are 1) enlarging the linguistic assessment items to other language pairs other than English-to-Korean, 2) semi-automatically collecting the source text which is targeted for evaluation, 3) extending the research to machine interpreting with speech data, 4) including the assessment items that human translator considers.

▸ Keywords: machine translation, neural machine translation, taxonomy, test suites, linguistic assessment item
▸ 주제어: 기계번역, 신경망 기계번역, 분류체계, 평가세트, 평가항목

최승권
한국전자통신연구원 언어지능연구실 책임연구원
choisk@etri.re.kr
관심분야: 기계번역, 자동통역, 대화처리, 코퍼스 가공

한지은
한국외국어대학교 ELLT학과 학생
faithjieunhan@naver.com
관심분야: 기계번역, 자동통역, 토픽모델링, 코퍼스 가공

최규현
한국전자통신연구원 언어지능연구실 연수생
choko93@etri.re.kr
관심분야: 기계번역, 자동통역

김영길
한국전자통신연구원 언어지능연구실 책임연구원
kimyk@etri.re.kr
관심분야: 기계번역, 자동통역, 인공지능

논문투고일: 2020년 11월 8일
심사완료일: 2020년 11월 26일
게재확정일: 2020년 11월 30일