자동통번역 시스템의 언어 현상별 자동 평가

최승권*, 최규현**, 김영길*
*한국전자통신연구원 언어지능연구실
**과학기술연합대학원대학교 컴퓨터소프트웨어
e-mail : {choisk, choko93, kimyk}@etri.re.kr

Automatic Evaluation of Speech and Machine Translation Systems by Linguistic Test Points

Sung-Kwon Choi*, Gyu-Hyun Choi**, Young-Gil Kim*
*Language Intelligence Research Section, ETRI
**Dept. of Computer Software, UST

요 약

자동통번역의 성능을 평가하는데 가장 잘 알려진 자동평가 기술은 BLEU 이다. 그러나 BLEU 로는 자동통번역 결과의 어느 부분이 강점이고 약점인지를 파악할 수 없다. 본 논문에서는 자동통번역 시스템의 언어 현상별 자동평가 방법을 소개하고자 한다. 언어 현상별 자동평가 방법은 BLEU 가 제시하지 못하는 언어 현상별 자동평가가 가능하며 개발자로 하여금 해당 자동통번역 시스템의 언어 현상별 강점과 약점을 직관적으로 파악할 수 있도록 한다. 언어 현상별 정확도 측정은 Google 과 Naver Papago 를 대상으로 실시하였다. 정확률이 40%이하를 약점이라고 간주할 때, Google 영한 자동번역기의 약점은 스타일(32.50%)번역이었으며, Google 영한 자동통역기의 약점은 음성(30.00%)인식, 담화(30.00%)처리였다. Google 한영 자동번역기 약점은 구문(34.00%)분석, 모호성(27.50%)해소, 스타일(20.00%)번역이었으며, Google 한영 자동통역기 약점은 담화(30.00%)처리였다. Papago 영한 자동번역기는 대부분 정확률이 55% 이상이었으며 Papago 영한 자동통역기의 약점은 담화(30.00%)처리였다. 또한 Papago 한영 자동번역기의 약점은 구문(38.00%)분석, 모호성(32.50%)해소, 스타일(20.00%)번역이었으며, Google 한영 자동통역기 약점은 담화(20.00%)처리였다. 언어 현상별 자동평가의 궁극적인 목표는 자동통번역기의 다양한 약점을 찾아내어 약점과 관련된 targeted corpus 를 반자동 수집 및 구축하고 재학습을 하여 자동통번역기의 성능을 점증적으로 향상시키는 것이다. [1]

1. 서론

현재 딥러닝을 기반으로 음성인식 및 자동번역의 성능이 비약적으로 발전하고 있어 자동통번역 시스템도 그 발전 속도가 무척 빠르다. 자동통번역의 성능을 평가하는 기술로 가장 잘 알려진 것이 BLEU 이다 [1]. BLEU 의 장점은 평가하고자 하는 원문에 대해 정답인 번역문(Reference)이 존재하면 시스템을 자동으로 평가할 수 있다는 것이다. 즉, 원문에 대한 자동번역 결과를 정답인 번역문(Reference)과의 n-gram 유사도를 계산하여 자동으로 평가를 수행하는 것이다. 반면에 BLEU 의 단점은 평가 점수만 보고서는 해당 시스템의 문제점이 무엇인지를 파악할 수 없으며, 평가에 사용한 정답인 번역문에 의존적인 결과를 내린다는 것이다 [2]. 따라서 개발자나 사용자들은 BLEU 의 평가 점수만 가지고는 자동통번역 시스템의 어느 부분이 장점이고 어느 부분이 단점인지를 파악하기 어렵다.

따라서 본 논문에서는 자동통번역 시스템의 장점과 단점을 자동으로 파악할 수 있는 방법인 언어 현상별 자동평가 방법을 제안하고자 한다. 언어 현상별 자동평가 방법에 의하면 BLEU 에서 제시하지 못하는 언어 현상별 자동평가가 가능하며 개발자 관점에서 자동통번역 시스템의 언어 현상별 장점과 단점을 직관적으로 파악하여 자동통번역 시스템의 단점을 개선할 수 있다.

2. 관련 연구

본 논문의 언어 현상별 평가 방법과 같이 자동번역 시스템의 장단점을 파악하기 위해 평가셋(Test suites)을 구축하는 연구가 있었다. Bentivogli [3]는형태소, 어휘, 어순과 관련된 IWSLT 2015 의 English-German 평가셋을 만들어 PBMT(Phrase-Based Machine Translation)와 NMT(Neural Machine Translation)의 장단점을 평가하였다. Isabelle [4]은 영불 NMT 의 구조적인 차이를 분

[1] 이 논문은 2019 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)

류하고 장단점을 밝혀내기 위해 108 개의 영-불 평가셋 문장을 구축하고 Yes/No 질문에 따라 평가하였다. Guillou [5]은 DiscoMT 2015 shared task 에서 주어 위치의 대명사 it 과 they 의 영불 자동 번역에 대한 평가를 실시하기 위해 평가셋을 구축하였다. Brussel [6]은 NMT, PBMT, RBMT(Rule-based MT)를 비교하기 위해 자동번역 오류 평가셋을 구축하였다.

또한 본 논문의 언어 현상별 자동평가 방법과 같이 자동으로 평가하는 방법에 대한 연구가 있었다. 독-영 평가셋에서 평가대상이 되는 평가 현상을 정규식 표현(Regular expression)으로 기술하여 자동으로 평가하는 방법이었다[7]

기존 논문과 달리 본 논문에서는 자동번역 언어 현상뿐 아니라 자동통역과 관련된 언어 현상들도 평가할 수 있다는 것이 차별화된 점이다.

3. 언어 현상별 평가셋 구성

3.1. 평가셋의 구조

언어 현상별 자동평가용 평가셋은 <원문> <원문어휘> <평가어휘>로 구성된다. <원문>은 평가할 언어현상별 출발 언어의 문장을 말하며, <원문어휘>는 <원문>에 포함된 언어 현상별 평가어휘를 말한다. <평가어휘>는 <원문어휘>에 대응되는 정답 대역 표현을 말한다. 언어 현상별 <원문>과 <원문어휘>는 수동으로 구축하였으며, <평가어휘>는 인터넷 번역사전으로부터 수집하였다. 평가셋은 다음과 같은 다양한 구조를 가진다.

1) <원문> <원문어휘> <평가어휘>
평가셋의 기본 구조는 1 개의 <원문>, 1 개의 <원문어휘>, 1 개의 <평가어휘>로 구성된다.
예) <In these 3 years, 2.5 billion pencils were sold> <2.5 billion> <25 억>

2) <원문> <원문어휘> <평가어휘 1>… <평가어휘 n>
평가셋은 <원문어휘>에 대응되는 2 개 이상의 <평가어휘>로 구성될 수 있다.
예) <I could not but get angry> <could not but> <지 않고는 있을 수 없었> <지 않고 있을 수 없었> <지 않을 수 없었>

3) <원문> <원문어휘> <#평가어휘 1>… <#평가어휘 n>
<#평가어휘>는 자동 통번역한 결과 중 <평가어휘>로 간주될 수 있는 <평가어휘>를 말한다.
예) <I could not but get angry> <could not but> <지 않고는 있을 수 없었> <지 않고 있을 수 없었> <지 않을 수 없었><#수 밖에 없었>

4) <원문> <원문어휘> <평가어휘 11//평가어휘 12>… <평가어휘 n1//평가어휘 n2>
<평가어휘 11//평가어휘 12>는 <원문어휘>에 대응되는 <평가어휘>가 <평가어휘 11>과 <평가어휘 12>로 분리되어 나타날 수 있음을 말한다.
예) <None of the books were interesting> <none of the books> <어느 책도//지 않>, <책들 중 어느 것도//지 않>

5) <원문> <원문어휘> <~평가어휘>
<~평가어휘>는 <원문어휘>에 대응되는 <평가어휘>가 자동통번역문에 나타나지 말아야 한다는 것을 의미한다. 즉 '~'은 논리적 결합어 중에서 논리적 부정(negation)을 표현한다.
예) <It was September 17> <It> <~그것>

3.2. 언어 현상별 평가셋

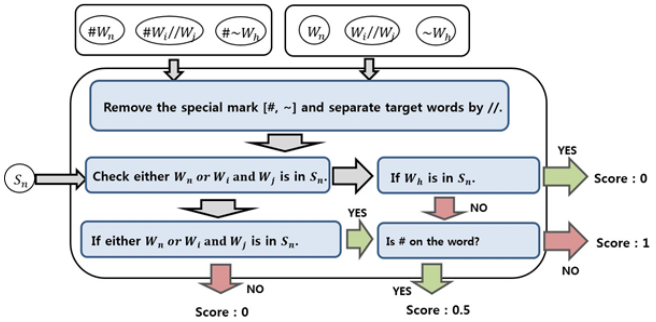
언어 현상별 평가셋은 자동번역과 자동통역의 평가셋으로 구분하여 구축하였다. 자동번역을 위한 언어 현상별 평가셋의 언어 현상 항목수는 58 개이고 해당 문장수는 630 문장이다 [8]. 자동통역을 위한 언어 현상별 평가셋의 언어 현상 항목수는 6 개이고 해당 문장수는 150 문장이다.

<표 1> 언어 현상의 항목수 및 문장수

구분	분류 1	분류 2	#항목	#문장
자동번역	품사	관사, 고유명사, 인칭대명사, 형용사, 부사, 전치사, 동사, 조동사, 관계대명사, 관계부사, 접속사, 기호, 수사 등	29	290
	구문	부정사 구문, 분사 구문, 동명사, 관용표현 등	10	100
	문장	문장종류, 부정문, 비교구문, 가정법, 수동태, 화법, 삽입, 생략도치, 병렬, 시제 등	14	140
	모호성	품사 모호성, 구조적 모호성	2	40
	공기관계	공기관계	1	20
	다의어	다의어	1	20
	스타일	자연스러운 번역표현	1	20
자동통역	음성	Spoken표현	1	100
	통역	장문	1	10
		Named entity	1	10
		통역단위	2	20
	담화	Cohesion	1	10
			64	780

4. 언어 현상별 자동평가 방법

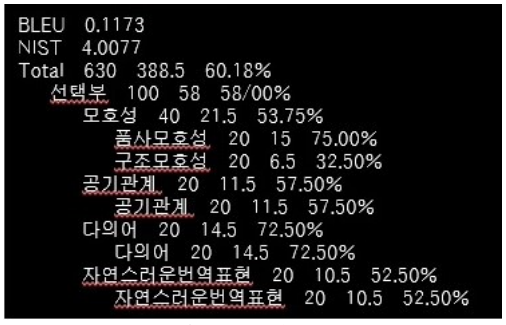
언어 현상별 자동평가 구성도는 다음과 같다.



(그림 1) 언어 현상별 자동평가 방법

그림에서 Sn 은 <원문>의 자동번역 결과를 말하며, Wn, Wi, Wj, Wh 는 <평가어휘>를 의미한다. 프로그램은 '#'의 유무를 판단하고 나서 '~'와 '//'가 평가어휘에 표시되어 있는지 확인한다. <평가어휘>가 자동번역 결과에 존재하면 1 점을 부여하고 존재하지 않으면 0 점을 부여한다. <#평가어휘>가 자동번역 결과에 존재하면 0.5 점을 부여하고 존재하지 않으면 0 점을 부여한다. <~평가어휘>가 자동번역 결과에 존재하면 0 점을 부여하고 존재하지 않으면 1 점을 부여한다. <평가어휘 1//평가어휘 2>가 자동번역 결과에 존재하면 1 점을 부여하고 존재하지 않으면 0 점을 부여한다.

이상의 언어현상별 자동평가 방법에 의해 만들어진 결과는 다음과 같은 모습으로 프린트된다.



(그림 2) 언어 현상별 자동평가 결과

5. 실험

실험에 사용된 자동통번역 시스템은 Google translator 와 Naver 의 Papago 였다. 언어쌍은 영한과 한영에 대해 이루어졌다. 자동통역 시스템의 실험에 사용한 음성 데이터는 원본 동영상 파일을 2~3 분 단위로 자른 동영상클립의 음성인식 및 동시통역 자동분절 결과를 입력으로 간주하였다.

<표 2> 언어 현상별 정확률

	영한		한영	
	Google	Papago	Google	Papago
Bleu	0.0976	0.2303	0.3662	0.3999
Nist	3.9359	5.4058	7.3951	7.5919
전체	42.67%	57.03%	46.92%	49.06%
자동번역	51.35%	67.06%	39.84%	44.13%
●품사	53.28%	69.83%	42.76%	46.90%
●구문	41.50%	57.50%	34.00%	38.00%
●문장	55.36%	71.07%	41.43%	46.43%
●모호성	48.75%	63.75%	27.50%	32.50%
●공기관계	50.00%	55.00%	40.00%	40.00%
●다의어	70.00%	77.50%	60.00%	70.00%
●스타일	32.50%	55.00%	20.00%	20.00%
자동통역	34.00%	47.00%	54.00%	54.00%
●음성	30.00%	47.50%	62.00%	62.00%
●통역	45.00%	50.00%	40.00%	42.50%
●담화	30.00%	30.00%	30.00%	20.00%

표 2 의 언어현상별 정확률에 따르면 Google 과 Papago 의 영한 언어현상별 자동통번역 정확률은 42.67% 대 57.03%로 Papago 의 정확률이 14.36% 높았다. Google 과 Papago 의 한영 자동통번역 언어현상별 정확률은 46.92% 대 49.06%로 Papago 의 정확률이 2.14% 높았다. 표 2 는 Google 과 Papago 의 자동번역과 자동통역 장단점도 일목요연하게 보여주고 있다. 정확률이 40%이하를 약점이라고 간주할 때, Google 영한 자동번역기의 약점은 스타일(32.50%)번역이었으며, Google 영한 자동통역기의 약점은 음성(30.00%)인식, 담화(30.00%)처리였다. Google 한영 자동번역기 약점은 구문(34.00%)분석, 모호성(27.50%)해소, 스타일(20.00%)번역이었으며, Google 한영 자동통역기 약점은 담화(30.00%)처리였다. Papago 영한 자동번역기는 대부분 정확률이 55% 이상이었으며 Papago 영한 자동통역기의 약점은 담화(30.00%)처리였다. 또한 Papago 한영 자동번역기의 약점은 구문(38.00%)분석, 모호성(32.50%)해소, 스타일(20.00%)번역이었으며, Google 한영 자동통역기 약점은 담화(20.00%)처리였다.

6. 결론

본 논문에서는 언어 현상별 자동평가 방법을 소개하였다. 자동통번역 시스템의 언어 현상별 자동평가 방법은 다음과 같이 이루어졌다. 1) 원문을 자동번역 후, <평가어휘>가 자동번역문에 있으면 1 점을 부여한다. 2) 원문을 자동번역 후, <#평가어휘>가 자동번역문에 있으면 0.5 점을 부여한다. 3) 원문을 자동번역 후, <평가어휘>가 자동번역문에 없으면 0 점을 부여한다. 이러한 언어 현상별 자동평가 방법에 따라 2 개의 자동통번역기를 평가하였으며 언어 현상별 장단점을 파악할 수 있었다. 언어 현상별 자동평가 방법의 궁극적인 목표는 자동통번역기의 다양한 약점을 찾아내어 약점과 관련된 targeted corpus 를 반자동 수집 및 구축하고 재학습을 하여 자동통번역기의 성능을 점증적으로 향상시키는 것이다.

참고문헌

[1] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a method for automatic evaluation of machine translation”. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, US., 311–318, 2002.

[2]Lommel, Arle, Aljoscha Burchardt, Maja Popovic, Kim Harris, Eleftherios Avramidis, Hans Uszkoreit. “Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data”. In Proceedings of the 17th Annual Conference of the European Association for Machine Translation, 165-172, 2014.

[3] Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. “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, 2016.

[4] Isabelle, Pierre, Colin Cherry and George Foster. “A Challenge Set Approach to Evaluating Machine Translation”. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2486–2496, 2017.

[5] Guillou, Liane and Christian Hardmeier. “PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation”. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 636-643. 2016.

[6] Brussel, Laura Van, Arda Tezcan, Lieve. “A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch”. In Proceedings of International Conference on Language Resources and Evaluation (LREC), 3799-3804, 2018.

[7] Macketanz, Viven, Renlong Ai, Aljoscha Burchardt and Hans Uszkoreit. “TQ-AutoTest-A Semi-Automatic Test Suite for (Machine) Translation Quality”. In Proceedings of International Conference on Language Resources and Evaluation (LREC), 886-892, 2018.

[8] Choi, Sung-Kwon, Gyu-Hyeun Choi, and Youngkil Kim. “Automatic Evaluation of English-to-Korean and Korean-to-English Neural Machine Translation Systems by Linguistic Test Points”. In Proceedings of PACLIC, 2018.