영한 논문 번역시스템의 수동 평가와 자동 평가의 관계

최승권*, 황영숙**, 김영길*
*한국전자통신연구원 언어처리연구팀
**SK Telecom 미래사업개발 3 팀
e-mail : choisk@etri.re.kr

A Study to Relation between Human Judgment and Automatic Evaluation in English-Korean Scientific Paper MT System

Sung-Kwon Choi*, Young-Sook Hwang**, Young-Gil Kim*
*Natural Language Processing Research Team, Electronics and Telecommunications Research Institute
**Future Business Development Division, SK Telecom

요 약

본 논문은 영한 과학기술 논문 자동번역 시스템을 대상으로 이루어진 수동 평가와 자동 평가 사이의 상관 관계를 밝힘으로써 수동 평가와 자동 평가 중에 한쪽의 방법에 의해서 평가가 이루어지더라도 다른 쪽의 수치를 파악할 수 있도록 하는데 목표가 있다. 본 논문에서 수행한 수동 평가는 5 인의 전문 번역가가 5 회에 걸쳐 평가한 결과이며, 자동 평가는 영어 원문 1,000 문장에 대한 8 인이 번역한 8,000 문장의 정답문(References)과 자동번역 결과를 어절 단위와 형태소 단위로 N-gram 비교를 통해 평가된 결과이다. 본 논문에서 도출된 식은 사용하는 평가 집합과 대상 번역 시스템 별로 자동 평가와 수동 평가 간의 상관 계수를 만들어내고 수동 번역률을 구하는 식을 동일하게 적용한다면 시스템의 자동 평가 결과로부터 성능을 직관적으로 해석하는데 상당히 도움이 될 것이다.

1. 서론

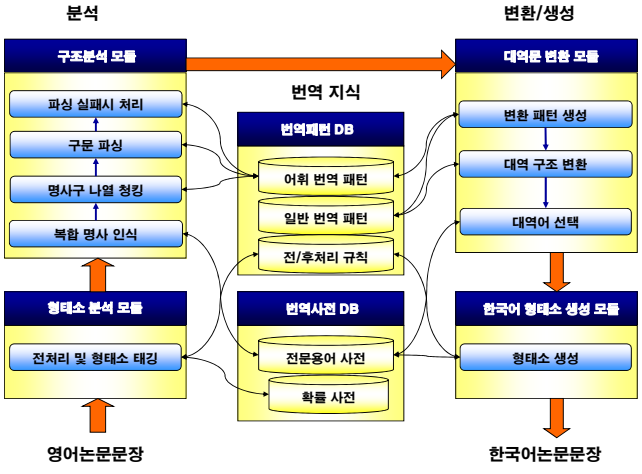
한국전자통신연구원(이하 ETRI)에서는 2005 년부터 2006 년까지 특허문서를 대상으로 영한 특허문서 자동번역 시스템을 개발하였다. 이 특허문서 자동번역 시스템은 상용화에 성공하여 산업자원부 특허지원센터에서 2007 년부터 중소기업을 대상으로 특허문서 영한 자동번역 서비스를 제공하고 있다.[1][2] 영한 특허문서 자동번역기를 성공적으로 개발한 이후, ETRI 에서는 2007 년부터 영어 과학기술 논문을 대상으로 영한 과학기술논문 자동번역 시스템을 개발하였다.

영한 번역 시스템에 대한 평가는 2006 년도까지 전문 번역가에 의한 수동 평가로만 이루어졌으나, 비싼 평가 비용, 긴 평가 기간, 평가자의 주관성과 같은 문제점 때문에 2007 년도에는 자동평가로 대변될 수 있는 BLEU [3] 점수에 의한 자동 평가가 수동 평가와 동시에 실시되었다. BLEU 에 의한 자동 평가는 평가 시간 절약, 적은 평가 비용, 튜닝에 큰 도움이 되었다. 그러나 아쉬운 점은 BLEU 점수가 직관적으로 무엇을 의미하며, 수동 평가와는 어떤 연관성이 있는지를 알 수 없다는 것이었다.

이에 본 논문에서는 영한 과학기술논문 번역 시스템을 대상으로 이루어진 수동 평가와 자동 평가의 결과를 가지고 둘 간의 상관 관계를 밝히는 것을 목표로 하고자 한다.

2. 시스템 구성도

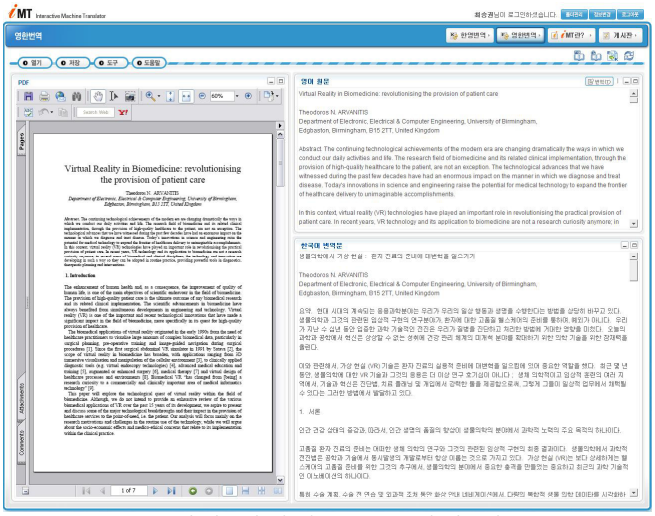
영한 과학기술논문 번역 시스템은 번역 방법론과 관련하여 패턴기반 자동번역 시스템[4]에 속하며, 그 전체적인 번역절차는 형태소 분석, 구조분석, 대역문 변환, 한국어 형태소 생성의 전형적인 분석-변환-생성의 형태를 갖추고 있다. 다음의 그림은 영한 과학기술논문 번역 시스템의 구성도를 개략적으로 기술한 것이다.



(그림 1) 시스템 구성도

위와 같은 구조를 가지는 영한 과학기술논문 번역 시

스템은 PDF 영어 논문을 TEXT 로 변환한 후, 변환된 TEXT 를 한국어 TEXT 로 번역하는 UI 를 가지고 있다. 다음의 그림이 영한 과학기술논문 번역 시스템의 번역 UI 를 보여준다.



(그림 2) 영한 과학기술논문 번역 시스템 UI

3. 수동 평가 방법

영한 번역 시스템에 대한 수동 평가는 전문 번역가가 평가 지침에 따라서 수작업으로 평가하는 것을 말한다. 영한 과학기술논문 번역 시스템의 수동 평가 방법은 다음과 같이 이루어졌었다.

수동 평가 문장: 자동 평가용 정답문 1,000 문장에서 임의로 400 문장을 선정함.

수동 평가 점수 기준: LREC(Conference on Language Resources and Evaluation)에서 사용하는 점수부여 기준표를 사용하였으며, 이 점수 기준은 원문의 의미를 얼마나 충실히 번역문에 전달하는가의 측면이 강조된 점수 부여 기준이다.

<표 1> 점수 부여 기준

4 점	All meaning expressed in the source fragment appears in the translation fragment
3 점	Most of the source fragment meaning is expressed in the translation fragment
2 점	Much of the source fragment meaning is expressed in the translation fragment
1 점	Little of the source fragment meaning is expressed in the translation fragment
0 점	None of the meaning expressed in the source fragment is expressed in the translation fragment

수동 평가 방법: 5 인의 번역가에게 '수동 평가 점수 부여 기준'을 교육한 후, 평가를 실시하고 문장당 최고/최저 점수를 제외한 3 인의 평균값으로 평가를 함.

번역률(%): $(\sum_{i=1}^{n}(\sum_{j=1}^{3}(median\_score_j/4))/3)/n \times 100.0$,

여기서 n 은 추출된 평가문의 수이며, $median\_score_j$ 는 j 번째 전문 번역가에 의해 평가된 점수가 5 개의 점수에서 최고, 최저 점수가 아닌 점수를 말한다.

4. 자동 평가 방법

자동 평가 방법은 개발중인 시스템의 성능 추이를 분석해가면서 시스템의 개발 진행 방향을 가늠하고 결정하는데 유용하게 사용될 수 있다. 영한 과학기술논문 번역 시스템의 자동 평가 방법은 보편적으로 널리 사용되는 BLEU[3]를 사용하였으며, 시스템 평가를 위해 다음과 같이 정답 집합(Reference Set)을 구축하였다:

정답 집합: 과학기술논문의 평균 문장수를 기초로 하여 5 개 분야(기계, 전기전자, 화학일반, 의료위생, 컴퓨터)에서 자동으로 영어 문장 1,000 문장을 수집하고, 각각의 문장에 대해 8 인의 전문 번역가가 한국어로 번역을 하였다. 1,000 문장과 관련해 정답문의 수를 더 많이 확보하면 보다 정확한 BLEU 평가 결과를 얻을 수 있겠지만[5], 번역 비용과 관련해 단일 정답문을 대량으로 만들어 자동평가를 하여도 다중 정답문의 평가와 유사할 수 있다.[3] 정답문(References)의 영어 원문 단어수와 문장당 평균 단어수는 다음과 같았다:

<표 2> 정답문의 영어 원문 문장당 평균 단어수

분야	문장수	단어수	평균단어수
기계	200	3,638	18.19
전기전자	200	3,559	17.80
화학일반	200	3,849	19.25
의료위생	200	3,944	19.72
컴퓨터	200	3,773	18.87
계	1,000	18,763	18.76

정답문들 사이의 자동 평가: 전문 번역가가 번역한 한국어 문장들 사이를 자동 평가한 결과를 말함.

자동번역 결과와 정답문과의 자동 평가: 자동 번역 결과와 정답문과의 상호 자동 평가한 결과를 말함.

5. 수동 평가와 자동 평가와의 관계

5.1. 수동 평가 실험

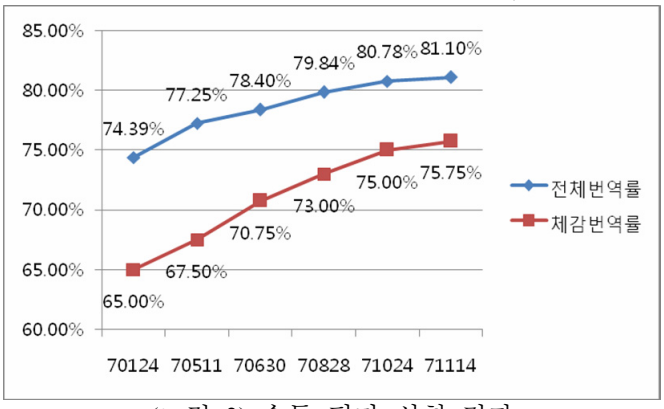
수동 평가 실험은 총 6 회에 걸쳐 실시되었다. 수동 평가 실험에 사용된 샘플 문장은 앞서 3 절에서 기술된 바와 같이 정답문에서 임의로 추출한 영어 원문 400 문장이었다. 이 400 문장은 6 회에 걸친 수동 평가 실험에서 동일하게 적용되었으며, 평가자는 매번 갱신되는 자동번역 시스템에 의한 번역 결과를 영어 원문과 비교하면서 평가를 하였다.

수동 평가 실험에 사용된 400 문장의 단어수와 문장당 평균 단어수는 다음과 같았다:

<표 3> 수동평가 집합의 문장당 평균 단어수

문장수	단어수	문장당 평균 단어수
400	7,332	18.33

6 회에 걸친 수동 평가 결과는 다음의 그림과 같았다:



(그림 3) 수동 평가 실험 결과

위의 그림에서 전체 번역률은 점수 부여 기준 0~4 전체에 대한 번역률을 말하며, 체감 번역률은 점수 부여 기준에서 3 점 이상의 문장에 대한 번역률을 말한다. 체감 번역률은 개발자 기준보다는 일반 사용자 입장에서 번역 결과가 이해가 되느냐 안되느냐의 관점에서 본 번역률이라 할 수 있다. 위의 그림에 따르면 영한 과학기술논문 번역 시스템의 전체 번역률은 2007 년 1 월 24 일 74.39%에서 2007 년 11 월 14 일 81.10%로 6.71% 향상되었으며, 체감 번역률은 65%에서 75.75%로 10.75% 개선되었다.

5.2. 자동 평가 실험

자동 평가 실험은 모듈의 튜닝, 사전 또는 패턴 수에 변화가 있을 때마다 실시되었다.

5.2.1. 정답문들 사이의 자동평가

전문 번역가는 다른 전문 번역가의 번역 결과에 대해 어떻게 평가하는지를 자동 평가하여 보았다. 정답문을 대상으로 자동 평가한 결과는 다음과 같았다:

<표 4> 정답문의 자동 평가 결과(BLEU)

BLEU-EO1	BLEU-MO1
0.435	0.661525

위에서 '-EO'은 정답문의 한국어 번역문을 어절 형태에서 자동 평가한 것을 말하며, '-MO'는 정답문의 한국어 번역문을 형태소 단위에서 자동 평가한 것을 말한다. 따라서 위에서의 BLEU-EO1 값인 0.435 와 BLEU-MO1 값인 0.661525 는 각각 어절 단위 자동 평가와 형태소 단위 자동 평가에서 100%의 번역률이라고 간주할 수 있을 것이다.

5.2.2. 자동번역 결과와 정답문과의 자동평가

자동 번역된 결과와 정답문 사이의 자동평가도 어절 단위와 형태소 단위로 구분하여 평가하였다. 그 결과는 다음과 같았다:

<표 5> 정답문의 자동 평가 결과

	BLEU-EO2	BLEU-MO2
20070511	0.273	0.4946
20070613	0.2875	0.5104
20070622	0.2928	0.5176
20070628	0.2944	0.5189
20070630	0.2918	0.5152
20070711	0.2952	0.5168
20070723	0.2966	0.5187
20070724	0.2958	0.519
20070828	0.2934	0.5169
20071001	0.2953	0.5165
20071012	0.2892	0.5067
20071024	0.3004	0.5098
20071114	0.3081	0.5185

5.3. 수동 평가와 자동 평가와의 관계

인간에 의한 수동 평가 결과가 자동 평가 결과와 매우 유사하다는 실험 결과는 수동 평가와 BLEU 에서 [3][6], 수동평가와 METOR(단어 일치 뿐만 아니라 동의어 일치, 형태소 분석된 어휘 일치를 사용한다는 점에서 BLEU 평가와 차이가 있다.)에서[7] 제시된 바 있다. 영한 자동번역 시스템에서의 5 회에 걸쳐 실시된 수동 평가와 동일한 시점에 이루어진 자동 평가와의 상관 관계를 나타내는 결과는 다음의 표와 같았다:

<표 6> 수동 평가와 자동 평가 표

날짜	BLEU-MO2	BLEU-MO2/BLEU-MO1	수동번역률 (%)	수동번역률 /(BLEU-MO2/BLEU-MO1)
070511	0.4946	74.77	77.25	1.033168
070630	0.5152	77.88	78.40	1.006677
070828	0.5169	78.14	79.84	1.021756
071024	0.5098	77.06	80.78	1.048274
071114	0.5185	78.38	81.10	1.034703
날짜	BLEU-EO2	BLEU-EO2/BLEU-EO1	수동번역률 (%)	수동번역률 /(BLEU-EO2/BLEU-EO1)
070511	0.2730	62.76	77.25	1.230880
070630	0.2918	67.08	78.40	1.168754
070828	0.2934	67.45	79.84	1.183692
071024	0.3004	69.06	80.78	1.169708
071114	0.3081	70.83	81.10	1.144995

또한 주어진 평가 집합을 대상으로 번역 시스템의 성능 변화 추이에 따른 수동 평가 결과와 자동 평가 결과 사이의 상관 관계를 분석한 결과, <표 7>과 같았다.

<표 7> 자동 평가와 수동 평가 사이의 상관계수

	BLEU-EO2	BLEU-MO2
상관계수	0.94	0.71

위의 도표는 어절 단위의 경우 수동 평가와 자동 평

가 결과가 0.94 로 둘간의 상관 관계가 아주 강한 것으로 나타났으며, 형태소 단위의 경우 어절 단위만큼 높지는 않으나 그래도 상관성이 높은 것으로 나타났다. 형태소 단위의 상관성이 높지 않았던 이유는 형태소 분석기의 오류가 평가결과에 영향을 미쳤기 때문이다.
이러한 수동 평가와 자동 평가의 밀접한 상관 관계를 토대로 자동 평가 결과에 의한 수동 평가를 예측할 수 있는 식을 기술하면 다음과 같다:

수동번역률
≒(BLEU-EO2/ BLEU-EO1 x 100) x 1.174051 ±α
≒(BLEU-MO2/ BLEU-MO1 x 100) x 1.029876 ±β

위의 식은 'BLEU-MO2/BLEU-MO1'와 'BLEU-MO2/BLEU-MO1'으로부터 최고, 최저 오차가 나는 경우를 제외한 각각의 3 개의 평균값을 구해서 얻은 1.02987 과 1.174051 를 이용하여 구한 식이다.
이와 같은 자동 평가 결과로부터 수동 평가 결과치(수동번역률)를 예측하기 위한 수식은 현재의 자동 평가를 위한 평가 집합과 시스템에 의존적인 것이다. 그러므로 일반화시켜 다른 시스템에 대해서도 동일하게 적용하기는 어렵다. 대신에 사용하는 평가 집합과 대상 번역 시스템 별로 자동 평가와 수동 평가 간의 상관 계수 혹은 상대 비율을 추정하여 수동 번역률의 근사식을 구하는 방법을 동일하게 적용한다면 시스템의 자동 평가 결과로부터 성능을 직관적으로 해석하는데 상당히 도움이 될 것이라 판단된다.

6. 결론

본 논문에서는 영한 과학기술논문 번역 시스템을 대상으로 전문 번역가에 의한 수동 평가와 정답문에 기반한 자동 평가 사이에 밀접한 상관 관계가 있음을 밝히고, 자동 평가의 BLEU 점수로부터 직관적으로 받아들일 수 있는 수동 평가의 점수를 얻기 위한 식을 기술하는 것이 목표였다.
본 논문에서 도출된 식은 사용하는 평가 집합과 대상 번역 시스템 별로 자동 평가와 수동 평가 간의 상관 계수를 만들어내고 수동 번역률을 구하는 식을 동일하게 적용한다면 시스템의 자동 평가 결과로부터 성능을 직관적으로 해석하는데 상당히 도움이 될 것이라 판단된다.
향후의 계획은 본 논문에서 제시한 수동 평가와 자동 평가의 관계식을 더욱 많은 실험에 의해 다듬어 개선하는 것일 것이다.

참고문헌

[1] 최승권, 권오욱, 이기영, 노윤형, 박상규. "도메인 특화 방법에 의한 영한 특허 자동번역 시스템의 구축", 한국정보과학회 논문지: 소프트웨어 및 응용. 제 34 권 제 2 호, 99-103 쪽, 2007.
[2] Oh-Woog Kwon, Sung-Kwon Choi, Ki-Young Lee, Yoon-Hyung Roh, Young-Gil Kim. "English-Korean Patent Translation System: FromTo-EK/PAT", In Machine Translation Summit XI: Workshop on Patent Translation, Copenhagen, Denmark, pp.1-7, 2007.
[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation". In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, USA, pp.311-318, 2002.
[4] Koichi Takeda. "Pattern-based machine translation", In Proceedings of the 16th conference on Computational linguistics – Volume 2, Copenhagen, Denmark, pp.1155-1158, 1996.
[5] Andrew Finch, Yasuhiro Akiba and Eiichiro Sumita. "How Does Automatic Machine Translation Evaluation Corelate With Human Scoring as the Number of Reference Translations Increases?", In Proceedings of LREC 2004, Vol 6, pp.2019-2022, 2004.
[6] Deborah Coughlin. "Correlating Automated and Human Assessments of Machine Translation Quality", In Proceedings of the Ninth Machine Translation Summit (MT Summit IX), New Orleans, USA, pp.63-70, 2003.
[7] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, USA, pp.65-72, 2005.