BLEU 를 활용한 단기 서술형 답안의 자동 채점

An Autonomous Assessment of a Short Essay Answer by Using the BLEU

조정현, Junghyun Cho*, 정현기, Hyunki Jung**, 박찬영, Chanyoung Park**, 김유섭, Yuseop Kim***

요약 본 논문에서는 단기 서술형 답안의 자동 채점을 위하여 기계 번역 자동 평가에서 널리 사용되는 BLEU(BiLingual Evaluation Understudy)를 활용한 방법을 제안한다. BLEU 는 기계가 번역한 것이 사람이 번역한 것과 비슷할수록 기계번역의 질이 좋을 것이다 라는 것을 가정하여 평가한다. 즉, 특정 문장을 여러 사람이 번역한 문장을 기계가 번역한 문장과 n-gram 방식으로 비교해 점수를 매기는 것이다. 이와 비슷하게 본 연구에서는 여러 개의 정답 문장과 학생의 답안 문장을 BLEU 와 같은 방식으로 상호 비교하여 학생의 답안을 채점하였다. 실험에서는 이러한 채점 방식의 정확도를 평가하기 위하여 사람이 채점한 점수와의 상관관계를 계산하였다.

Abstract We propose a method utilizing BLEU(BiLingual Evaluation Understudy), which is widely used in automatic evaluation of machine translations, for an autonomous assessment of a short essay answer. BLEU evaluates translations with an assumption that the translation by a machine is supposed to be more accurate as it is getting to be more similar to the translation by a human. BLEU scores the translation by comparing the n-grams of translations by a machine and humans. Similarly we score students answers by comparing to multiple reference answers with BLEU. In the experiment, we compute correlation coefficient values between scores of our system and human instructors.

핵심어: BLEU, assessment, answer, autonomous, n-gram

본 논문은 산업자원부의 지역혁신 인력양성사업(KOTEF)의 지원과 2006 년도 정부재원(교육인적자원부 학술연구조성사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2006-331-D00534)

*주저자 : 한림대학교 컴퓨터공학과 석사 e-mail: showcjh@hallym.ac.kr

**공동저자 : 한림대학교 컴퓨터공학과 석사 e-mail: mayapple@hallym.ac.kr

**공동저자 : 한림대학교 컴퓨터공학과 교수 e-mail: cypark@hallym.ac.kr

***교신저자 : 한림대학교 컴퓨터공학과 교수; e-mail: yskim@hallym.ac.kr

1. 서론

인터넷, 컴퓨터, 광랜의 보급으로 인해 e-learning 이 늘어나고 있다. 따라서 e-learning 에서의 시험등과 같은 평가 또한 많아지고 있다. 또한 e-learning 이외에도 교육환경에서의 서술형 주관식 문제 평가는 학습의 성과를 판단할 수 있는 중요한 방법으로 많이 사용하고 있다. 사람이 서술형 주관식 문제를 채점할 때는 객관성 유지가 쉽지 않고 또한 시간도 매우 오래 걸린다. 하지만 서술형 주관식 문제의 채점을 사람이 아닌 컴퓨터가 자동으로 할 수 있다면 채점의 객관성을 유지할 수 있고 채점의 시간도 매우 빨라서 실시간 피드백이 가능할 것이다.

문제가 객관식이나 단답형의 주관식은 자동으로 채점하기 어렵지 않을 것이다. 하지만 학생의 답과 모범답안의 색인어를 미리 구축한 유의어 사전에서 검색하여 나온 유의어와 색인어를 단순 비교한 방법[1], 의미커널을 구축하고 학생의 답과 모범답안을 벡터로 구성하여 이 답안간의 유사도를 의미커널을 통해 계산하는 방법[2], [2]와 유사하게 LSA(Latent Semantic Analysis)를 이용한 방법[3] 등과 같은 기존 서술형 주관식 문제 채점 연구에서 볼 수 있듯이 자연언어 처리를 하여 분석을 해야 하기 때문에 많은 어려움이 있다.

본 논문에서는 간단한 서술형 주관식 자동 채점을 빠르고 간단하게 하기 위해 BLEU[4]를 활용한 방법을 제안한다. 먼저 이전에 BLEU 를 활용하여 자동 채점을 시도한 연구가 있었으나[5] 이것은 영어나 스페인어와 같은 굴절어를 대상으로 하고 있으며, 또한 채점의 대상이 되는 답안의 길이가 30 내지 50 단어 정도로 매우 길었다.

따라서 본 연구에서는 교착어의 특성을 가지고 있는 한국어에 대하여 이 방식을 적용하여 그 결과를 분석하였으며, 또한 기존 연구와는 달리 1-2 문장으로 이루어진 짧은 길이의 답안의 채점에 이 방법을 새롭게 적용하였다.

BLEU 는 기계번역의 자동 평가에서 많이 사용하는 방법이다. 이 방법을 이용해 기계번역의 품질에 점수를 매기는 것이다. 사람의 번역과 기계의 번역을 비교하여 얼마나 사람의 번역에 가까운지를 보는 것이다. 마찬가지로 자동 채점에서는 정답과 학습자의 답을 비교해 얼마나 정답에 가까운지를 볼 수 있을 것이다. 2 장에서 BLEU 에 대해 설명하고 3 장에서는 BLEU 를 활용한 자동 채점 방법, 4 장에서는 실험 및 결과, 마지막 5 장은 결론 및 향후 연구에 대해 설명할 것이다.

2. BLEU

BLEU(BiLingual Evaluation Understudy)는 기계번역(Machine Translation) 시스템의 자동 평가 방법 중 하나이다. 또한 현재 많은 기계번역 연구에서 사용되고 있는 평가 방법이다. 이 BLEU 의 기본 전제 아이디어는 기계번역이 전문 번역인에 의한 번역과 비슷할수록 기계번역의 질은 좋을 것이라는 것이다. BLEU 는 전문 번역인이 번역한 뜻이 같은 여러 개의 번역을 하나의 코퍼스로 준비하여 이 코퍼스와 기계가 번역한 것을 비교한다. 비교는 n-gram(단어) 비교 매치 방법을 사용하여 수정된 P(Precision)을 구한다. 일반 precisions 의 식은 코퍼스의 n-gram 중에 기계번역의 n-gram 과 매치되는 총 횟수 / 기계번역의 총 n-gram 수 이지만 수정된 precision 의 식은 코퍼스의 n-gram 중에 기계번역의 n-gram 과 매치되는 최대값 / 기계번역의 총 n-gram 수 이다. 즉 중복된 n-gram 의 매치를 제거 한다는 것과 같은 뜻이다. n-gram 은 문장을 n 개의 어절로 나누는 것을 말한다. 즉, uni-gram 은 하나의 어절, bi-gram 은 두 어절로 나누는 것이다. BLEU 는 수정된 P 를 구하고 uni-gram, bi-gram, … , n-gram 에 대한 Pn(1)을 기하평균 값에 가중치 (2)를 지수승으로 곱한 식 (3)를 이용해 구한다.

$$P_n = \frac{\sum_{C\in\{Candidates\}} \sum_{n-gram\in C} Count_{clip}(n-gram)}{\sum_{C'\in\{Candidates\}} \sum_{n-gram'\in C'} Count(n-gram')} \quad (1)$$

$$\begin{cases} BP = 1 & if\ (c > r) \\ e^{(1-r/c)} & if\ (c \le r) \end{cases} \quad (2)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n \right) \quad (3)$$

식 (2)에서 c 는 기계번역 문장의 길이 이고 r 은 사람이 번역한 코퍼스 중에서 기계번역의 길이와 가장 비슷한 문장의 길이이다. c 가 r 보다 크면 가중치가 1 로 영향을 미치지 않으며 c 가 r 보다 작거나 같으면 가중치가 1-r/c 가 된다. 본 논문에서는 가중치 식 (2)는 사용하지 않고 식 (1)의 Pn 값에서 n-gram 을 uni-gram, bi-gram, tri-gram 각각을 계산한 값만 사용할 것이다.

3. BLEU 를 활용한 자동 채점 방법

주관식 문제는 고사성어 문제를 사용하였다. 즉, 고사성어를 보여 주고 그 뜻을 푸는 문제이다. 이 부분은 4 장 실험 및 결과에서 자세히 설명 할 것이다. 일단 문제는 총 30 문제이고 100 명의 학생이 답을 적었다. 한 문제당 모범 정답은 5 개씩 있다. 이 5 개가 하나의 코퍼스로 30 개의 문제 이므로 총 30 개의 코퍼스가 사용된다. 아래 그림 1 은 자동채점 방법을 구조화 한 것이다.

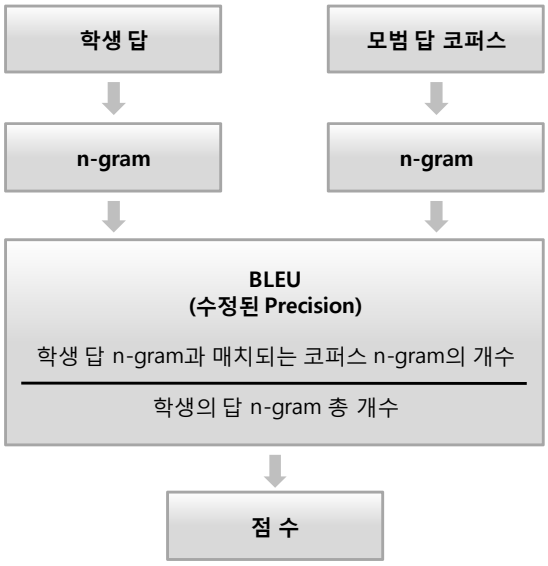

그림 1. BLEU 를 활용한 자동채점 방법

위 그림과 같이 코퍼스와 학생의 답을 n-gram 으로 나누고 그 둘을 수정된 precision 값을 구하는 방법으로 자동채점 하는 것이다. 먼저 uni-gram 으로 하나의 코퍼스와 학생의 답을 한 개씩 차례로 총 30 개를 비교한다. 코퍼스나 학생의 답에서 어절이 겹치는 것은 하나로 간주하고 비교하게 된다. 이는 수정된 precision 으로 계산 전에 미리 중복을 제거한 것이다. uni-gram 으로 나눈 학생의 답을 정답 코퍼스와 매칭하여 매치되는 카운트를 구하고 이 카운트를 학생 답의 uni-gram 개수로 나누게 되면 각 문제의 점수가 나오게 된다. 한문제당 점수는 0~1 로 한 학생당 총 30 점 만점으로 계산한다. n-gram 의 길이에 따른 점수 변화를 살펴보기 위해 uni-gram 과 같이 bi-gram, tri-gram 도 점수를 구한다.

4. 실험 및 결과

실험에 사용한 문제는 고사성어 문제이다. 고사성어를 제시하고 그 뜻을 풀이하는 문제이다. 문제는 총 30 문제로 각 학생당 30 문제를 100 명의 학생이 풀었다. 모범 답안은 한문제당 5 개로 그 뜻은 동일하다. 이렇게 각 문제당 5 개의 모범 답안으로 이루어진 총 30 개의 코퍼스를 준비했다. 실험은 uni-gram 과 bi-gram, tri-gram 으로 n-gram 길이에 따른 자동 채점 실험과 코퍼스의 크기, 즉 모범답안의 개수에 따른 실험을 하였다. 그리고 이 실험의 평가는 두 가지로 첫 번째는 자동 채점 점수와 사람이 수동 채점한 점수와의 상관관계를 구하고 두 번째는 학생의 답을 사람이 정답과 오답을 판별해 놓고 자동 채점한 점수를 0.01 에서 1.0 까지 Threshold 로 놓고 0.01 단위로 변경하면서 F-Measure 와 정확도(Accuracy)를 측정해 이 값이 가장 큰 Threshold 를 기준으로 잡고 정답과 오답을 자동으로 판별하였다. 그리고 이것을 증명하기 위해 4-fold cross-validation 을 이용해 F-Measure 와 정확도(Accuracy)의 평균을 측정하였다.

다음의 표 1 은 수동 채점(사람이 채점)한 100 명의 평균 점수, 표준편차와 5 개의 모범답안과 세 가지 n-gram 으로 채점한 평균 점수, 표준편차의 비교표 이다. 표 2 는 세 개의 n-gram 의 수동 채점과 자동 채점 점수의 상관계수와 모범답안의 개수에 따른 상관계수를 동시에 나타낸 표이다. 그림 2 는 수동 채점한 점수와 uni-gram 과 5 개의 모범답안으로 채점한 점수의 상관관계를 그래프로 보여주고 있다.

표 1. 수동 채점과 자동 채점의 점수 비교

채점	수동채점	uni-gram	bi-gram	tri-gram
평균점수	20.62	15.78	12.57	13.71
(표준편차)	(5.99)	(4.88)	(4.76)	(4.97)

표 2. 수동 채점과 자동 채점 점수의 상관계수

모범답안 개수	1 개	2 개	3 개	4 개	5 개
상관계수(uni-gram)	0.6682	0.7889	0.8600	0.8962	0.9364
상관계수(bi-gram)	0.5413	0.6789	0.7409	0.7883	0.8509
상관계수(tri-gram)	0.4969	0.6131	0.6834	0.7397	0.8054

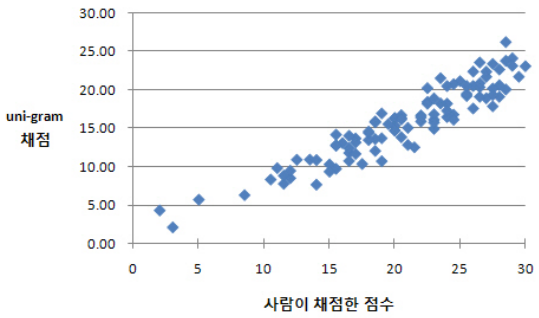

그림 2. 수동 채점과 uni-gram 채점의 상관관계

표 1 에서는 수동 채점과 uni-gram, bi-gram, tri-gram 자동 채점의 평균점수를 비교해 보여주고 있는데 uni-gram 의 평균 점수가 수동 채점 평균점수와 가장 근접한 것을 볼 수 있다. 표 2 는 모범 답안 개수와 n-gram 의 길이에 따른 수동 채점과 자동 채점 점수와의 상관계수로 uni-gram, bi-gram, tri-gram 모두에서 모범답안 개수가 늘어남에 따라 상관계수 역시 커짐을 볼 수 있다. 이는 모범답안의 개수가 5 개 까지는 그 개수가 늘어날수록 수동 채점과의 연관성이 높아진다고 볼 수 있다. 그 이유는 모범답안의 개수가 늘어나면 단어의 양이 많아지고 그에 따라 학생 답의 n-gram 이 모범답안의 n-gram 과 일치할 확률이 늘어나기 때문이다. 또한 n-gram 의 길이에 따른 수동 채점과의 상관계수는 uni-gram 채점이 가장 높게 나왔다. 표 1 에서 uni-gram 의 평균점수가 수동채점과 가장 근접한 결과와 같다고 볼 수 있다. 따라서 uni-gram 자동 채점이 가장 좋은 방법임을 알 수 있다. 모범답안이 5 개 일 때 uni-gram 채점 점수의 상관계수를 그래프로 보여주고 있는 그림 2 에서 보는 바와 같이 수동 채점과 uni-gram 채점의 상관관계가 강한 것을 볼 수 있다. 수동 채점한 점수와 자동 채점한 점수가 같지는 않지만 상관관계가 높아 두 점수의 연관성이 높다고 할 수 있다.

다음은 앞서 설명한 것과 같이 학생의 답을 사람이 정답과 오답을 판별해 놓고 자동 채점한 점수를 0.01 에서 1.0 까지 Threshold 로 놓고 0.01 단위로 변경하면서 F-Measure 와 정확도(Accuracy)를 측정해 이 값이 가장 큰 Threshold 를 기준으로 잡고 정답과 오답을 자동으로 판별하고 4-fold cross-validation 을 이용해 F-Measure 와 정확도(Accuracy)의 평균을 측정해 검증한 실험이다. 4-fold cross-validation 은 총 100 명의 학생이 쓴 답을 무작위로 25 개씩 추출하여 4 개 군으로 나누어 하나는 학습샘플로 나머지는 시험 샘플로 해서 총 4 번의 평균치를 내는 것이다. 표 3 과 표 5 는 Threshold 에 따른 F-Measure 와 정확도를 나타낸 것이며, 표 4 와 표 6 은 4-fold cross-validation 을 이용해 F-Measure 평균과 정확도 평균을 나타낸 것이다.

표 3. Threshold 에 따른 F-Measure

모범답안 개수	1 개	2 개	3 개	4 개	5 개
Threshold(uni-gram)	0.07	0.06	0.06	0.07	0.11
F-Measure	78.73%	84.22%	88.65%	89.23%	90.04%
Threshold(bi-gram)	0.06	0.07	0.07	0.07	0.07
F-Measure	65.52%	75.90%	81.93%	83.12%	86.23%
Threshold(tri-gram)	0.07	0.07	0.06	0.06	0.07
F-Measure	61.00%	73.17%	79.73%	81.32%	84.54%

표 4. 4-fold cross-validation 을 이용한 F-Measure 평균

모범답안 개수	1 개	2 개	3 개	4 개	5 개
F-Measure 평균 (uni-gram)	78.69%	84.15%	88.55%	89.10%	89.85%
F-Measure 평균 (bi-gram)	65.42%	75.78%	81.90%	83.07%	86.18%
F-Measure 평균 (tri-gram)	60.94%	73.10%	79.70%	81.23%	84.84%

표 5. Threshold 에 따른 정확도

모범답안 개수	1 개	2 개	3 개	4 개	5 개
Threshold(uni-gram)	0.07	0.06	0.06	0.11	0.11
정확도	71.20%	77.43%	82.93%	83.41%	84.73%
Threshold(bi-gram)	0.06	0.07	0.07	0.07	0.07
정확도	58.47%	68.40%	74.87%	76.17%	79.97%
Threshold(tri-gram)	0.07	0.07	0.06	0.06	0.07
정확도	54.57%	65.63%	72.37%	74.10%	77.90%

표 6. 4-fold cross-validation 을 이용한 정확도 평균

모범답안 개수	1 개	2 개	3 개	4 개	5 개
정확도 평균 (uni-gram)	71.16%	77.34%	82.70%	83.41%	84.49%
정확도 평균 (bi-gram)	58.38%	68.28%	74.83%	75.98%	79.87%
정확도 평균 (tri-gram)	54.51%	65.57%	72.33%	74.00%	77.82%

표 3, 표 4, 표 5, 표 6 에서 공통적으로 상관계수와 마찬가지로 모범답안이 5 개일 때 그리고 n-gram 의 길이가 uni-gram 일 때 F-Measure 와 정확도가 가장 높은 것을 볼 수 있었다. 표 3 과 표 5 에서는 자동 채점한 점수가 0.11 이상인 학생의 답을 정답이라고 한다면 F-Measure 는 90.04%, 정확도는 84.73%로 높은 수치를 보였다. 또한 이 수치는 표 4 와 표 6 에서의 수치가 90.04%, 84.49%로 상당히 유사하여 샘플에 따른 성능차이는 거의 없다고 볼 수 있다.

5. 결론 및 향후 연구

본 논문에서는 BLEU 를 활용한 간단한 서술형 주관식 자동 채점 방법을 제안하였다. 비록 자동 채점한 점수와 사람이 채점한 점수가 차이가 있지만 사람이 채점한 점수가 높으면 자동 채점 점수도 높았으며 반대로 낮으면 자동 채점 점수도 낮았다. 또한 자동 채점 점수로 정답과 오답을 자동으로 판별한 결과도 사람이 정답과 오답을 판별한 결과와 유사한 것을 실험을 통해 볼 수 있었다. 따라서 간단한 서술형 주관식 자동 채점의 한 방법으로 BLEU 를 이용하면 형태소 분석 등과 같은 복잡한 자연언어 처리를 하지 않고 간단히 자동 채점이 가능할 것이다. 이후 연구에서는 BLEU 방식을 개선하여 사람이 채점한 점수와 자동 채점 점수의 차이를 조금 더 근소하게 해야 할 것이다.

참고문헌

[1] 박희정, 강원석, “유의어 사전을 이용한 주관식 문제 채점 시스템 설계 및 구현”, 컴퓨터교육학회논문지, Vol.6 No3, pp. 207-216, 2003

[1] 오정석, 조우진, 김유섭, 이재영, “의미 커널에 기반한 주관식 문제 채점 시스템”, 한국정보기술학회논문지, 제 3 권 제 4 호, pp. 95~104, 2005

[3] 정상목, 한병래, 송기상, “LSA 를 이용한 서술형 주관식 평가 시스템의 설계 및 구현”, 한국정보교육학회논문지, Vol.9 No.2, pp. 289-298, 2005

[4] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of ACL, pp. 311~318, 2002.

[5] Diana P´erez, Enrique Alfonseca and Pilar Rodr´ıguez, “Application of the BLEU method for evaluating free-text answers in an e-learning environment”, In Proceedings of the Language Resources and Evaluation Conference, 2004