Weighted BLEU: 단어 가중치 기반 BLEU*

오진영○ 김정무 차정원
창원대학교
psycheojy@cwnu.ac.kr, gersanga@cwnu.ac.kr, jcha@cwnu.ac.kr

Weighted BLEU: N-Gram weighted BLEU

Jinyoung Oh○, Jeong-Moo Kim, Jeong-Won Cha
Changwon National University

요 약

우리는 다양한 범위의 문장 생성이 가능한 문제에서 생성된 문장의 품질을 평가하는 새로운 방법을 제안한다. 정답 문장에 있는 각 평가 단위에 [-2, 2] 사이의 가중치를 부가한다. 페르소나를 나타내는 문장을 생성하는 문제에서 단어 가중치 기반 BLEU는 인간의 평가와 유사한 결과를 보였다.

1. 서 론

많은 자연어처리 일감들에서 다양하고 사람이 읽을 수 있는 문장을 생성하는 것이 중요하다. 기계번역, 요약, 동의어구(paraphrase) 생성 그리고 이미지 캡셔닝 등의 일감에서 문장 생성이 필요하다. 이러한 분야에서 가장 먼저 해결해야 하는 것은 결과물에 대한 자동 평가이다. 왜냐하면 평가를 위한 결과물이 많아서 사람이 평가한다는 것은 현실적이지 않다.

기계번역을 비롯한 많은 문제에서 BLEU[1], METEOR[2] 등을 이용하여 자동 평가가 많이 개선되었다. 비록 BLEU가 많은 단점이 있지만[3] 그 값은 기계번역 분야에서 인간의 평가와 우수한 상관관계가 있음을 보였다[4,5,6,7].

그렇지만 기계번역 이외의 분야에서는 BLEU가 잘 적용되지 못했다. 따라서 이를 보완하는 평가 방법이 제안되었다. 동의어구(paraphrase) 생성 분야에서 다양성을 평가하기 위한 iBLEU가 있다[8]. iBLEU에서는 입력 문장과 동의어구 사이의 BLEU 값을 계산하여 BLEU 값을 줄인다. 이 방법에서는 동의어구에 맞게 매개변수를 조절할 수 있다. [9]에서는 tf-idf를 이용하여 BLEU에서 n-gram에 가중치를 부여하는 방법을 제안하였다. [10]에서는 평가할 문장에 사람이 [-1, 1] 사이의 점수를 부여하고 이를 가중치로 사용하는 방법을 제안하였다. 제안된 방법들은 BLEU가 포착하지 못했던 생성문의 우수한 점을 포착하기 위해서 제안되었다.

그러나 다음 예와 같이 페르소나가 포함된 문장 생성과 같은 분야에서는 문장 전체를 새롭게 생성하는 것이 아니라 문장의 일부분만이 변화한다 (문장 b).[1] 이러한 상황에서는 기존에 사용하던 평가 방법을 바로 적용하는 것은 적절하지 않다. 왜냐하면 '시픈지', '말해' 그리고 '주떼요'가 들어가지 않아도 높은 성능을 낼 수 있기 때문이다.

a) 무엇을 하고 싶으신지 정확히 말씀해 주세요.
b) 무엇을 하고 시픈지 정확히 말해 주떼요.

본 논문에서는 문장에서 핵심이 되는 부분을 직접 계산하는 방법을 제안한다. 이 방법은 중요한 부분이 포함된 문장을 생성하고자 할 경우의 평가 척도로써 유용하다.

2. 생성된 문장 평가

입력 문서 m에 의해서 생성된 문서를 c라고 하자. 또한 정답 문장 집합을 r이라고 하자. $r_{i,j}$는 i번째 문장에서 j번째 참조 문장이다.[2] c_i는 i번째 문장을 표기한다. 여기서 $i \in \{i,...,I\}$이다. 이러한 경우에 BLEU 값은 다음과 같이 계산된다.

$$BLEU = BP \cdot exp(\sum_n \log p_n) \quad (1)$$

$$BP = \begin{cases} 1 & \text{만약} R < C \text{일경우} \\ e^{(1-R/C)} & \text{아닌경우} \end{cases} \quad (2)$$

여기서 R은 참조 문장의 길이를 나타내고 C는 출력 문장의 길이를 나타낸다. 문서 단위의 p_n은 다음과 같다.[3]

$$p_n = \frac{\sum_i \sum_{g \in n-gram(c_i)} \max_j \{\min\{\#_g(c_i), \#_g(r_{i,j})\}\}}{\sum_i \sum_{g \in n-gram(c_i)} \#_g(c_i)} \quad (3)$$

* 이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2017R1D1A1B03033534)

1) 아이 페르소나가 적용된 문장 예.

2) 번역에서 j는 보통 5로 한다. 논 논문의 예에서는 1이다.

3) n은 $n-gram$을 나타낸다.

BLEU의 평가결과는 참조 문장의 수가 증가하면 사람의 평가 결과와 상호관계가 증가한다[11,12].

3. weighted BLEU

본 논문에서 제안하는 BLEU는 출력 문장에 반드시 포함되어야 하는 부분에 사람의 정량적인 가중치 $w_{i,j} \in [-2,2]$를 $r_{i,j}$에 부여하고 이를 사용하여 기존 BLEU를 확장한다. 새로운 BLEU는 식 (1), (2)를 그대로 사용한다. 다만 p_n을 다음과 같이 재정의 한다.

$$p_n = \frac{\sum_i \sum_{g \in n-gram(c_i)} \max_j \{\min\{w_g \times \#_g(c_i), w_g \times \#_g(r_{i,j})\}\}}{\sum_i \sum_{g \in n-gram(c_i)} w_g \times \#_g(c_i)} \quad (4)$$

여기서 w_g는 특정 n-gram의 가중치이다. 즉 n-gram내에 우리가 원하는 특정 부분이 있으면 가중치를 주는 것이다. 가중치가 없는 부분은 1.0으로 간주한다. n-gram내에 우리가 원하는 부분이 2개 이상 존재한다면 그 중에서 큰 값을 선택한다. 이 가중치는 사용자가 미리 지정한다. 앞의 예문의 경우에는 ('시픈지', 1.2), ('말해', 1.1), ('주떼요', 1.3)와 같이 가중치를 줄 수 있다. 또한 비문 경우는 ('가읫돈 는군요', -1.6), ('가이아 이 라면', -1.7), ('가집계 무서우', -1.6), ('간 는 것', -1.2)와 같이 줄 수 있다.

4. 실험 결과

우리는 (일반 문장, 페르소나 문장) 형식의 2,000 문장 쌍을 구축하였다. 평가를 위해서 RNN(Recurrent Neural Network)-RNN, CNN(Convolution Neural Network)-RNN, CNN-CNN 형식의 Seq2Seq 문장 생성기를 구축하고 그 결과를 다섯 사람이 [1, 5] 사이의 값으로 평가하여 평균값을 구하여 [0, 1]으로 사상하였다.

실험 결과는 표 1에 보인다. 사람이 평가한 것은 페르소나가 적절하게 나타난 것과 동시에 문장이 정문인지도 평가하였다.

표 1. weighted BLEU, BLEU, 사람평가의 상관관계

	BLEU	weighted BLEU	사람평가
RNN-RNN	0.632	0.561	0.102
CNN-RNN	0.641	0.570	0.095
CNN-CNN	0.663	0.613	0.110

표 2. BLEU에서 n-gram별 성능

	B@1	B@2	B@3	B@4
	wB@1	wB@2	wB@3	wB@4
CNN-CNN	0.764	0.688	0.626	0.575
	0.784	0.644	0.545	0.479

표 2는 n-gram별 성능을 비교하였다. 표에서 보듯이 n이 커질수록 값이 급격하게 떨어지는 것을 확인할 수 있다. 이것은 출력 문맥에서 비문이 많이 포함됨을 잘 측정해 준다고 보여진다.

5. 결 론

제안한 BLEU는 페르소나와 같이 문장의 일부분이 변하는 결과를 생성하는 문제에 대해서 표준 BLEU보다 사람의 평가와 높은 상관관계를 보인다. 평가를 위해서 초기에 지불해야 할 비용이 존재한다. 그렇지만 평가를 위해서 정답 문서를 만드는 비용에 비해서는 아주 미미하며 한번 구축하고 나면 시스템을 개발하는 중간에는 수정하지 않아도 된다. 새 BLEU는 생성된 문장에서 강조해야 하는 부분에 대해서 보다 더 잘 평가할 수 있기 때문에 기계번역이 아닌 이미지 캡셔닝, 문서 요약 등 보다 넓은 분야에 잘 적용할 수 있을 것으로 본다.

6. 참고 문헌

[1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proc. of ACL, pages 311-318.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with im- proved correlation with human judgments. In Proc. of ACL Workshop on Intrinsic and Extrinsic Evalua- tion Measures for Machine Translation and/or Sum- marization, pages 65-72.

[3] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In EACL, pages 249- 256.

[4] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proc. of HLT, pages 138-145.

[5] Deborah Coughlin. 2003. Correlating automated and human assessments of machine translation

quality. In Proc. of MT Summit IX, pages 63-70.
[6] Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In Proc. of EMNLP, pages 172-176.
[7] Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level ma- chine translation metrics. In Proc. of NAACL-HLT, pages 1183-1191.
[8] Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In ACL, pages 38-42.
[9] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In CVPR.
[10] Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, Bill Dolan, 2015, deltaBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets, In Proc. of ACL and IJCNLP, pages 445-450.
[11] M. Przybocki, K. Peterson, and S. Bronsart. 2008. Official results of the NIST 2008 "Metrics for MA- chine TRanslation" challenge (MetricsMATR08). http://nist.gov/speech/tests/metricsmatr/2008/results/.
[12] Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evalu- ation. In Proc. of HLT-NAACL, pages 162-171.