BLEU Paper Review

Reference: a Method for Automatic Evaluation of Machine Translation, Kishore et al.

1. Introduction

BLEU 논문은 2002년에 나왔는데, 당시 Machine Translation이 얼마나 좋은지에 대한 기준을 제시했다는 점에서 큰 의의를 가진다. citation도 굉장히 많은데, 그 이유는 아마 MT 성능 평가의 기준을 제시한 초기 논문이기 때문일 것이다.

저자들은 MT의 성능을 판단하는 가장 근본적인 기준으로, 전문적인 사람의 번역에 MT가 더 가까워 진다면, 더 좋을 것이라고 가정한다. 이러한 가정 하에, 전문적인 사람의 번역이 있어야 하고, 사람의 번역과 MT간의 근접도를 계산할 metric이 필요하다. 논문에서는 아래와 같이 표현한다.

- a numerical 'translation closeness' metric
- a corpus of good quality human reference translations

논문에서는 위에서 언급한 대로 사람의 번역과 MT를 비교하며 논의를 진행한다. MT가 만든 번역은 Candidate이고 사람이 만든 번역은 Reference이다.

2. The Baseline BLEU Metric

아래 예시를 살펴보자.

Candidate 1: It is a guide to action which ensures that the military always obeys the commands of the party.
Candidate 2: It is to insure the troops forever hearing the activity guidebook that party direct.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed the directions of the party.

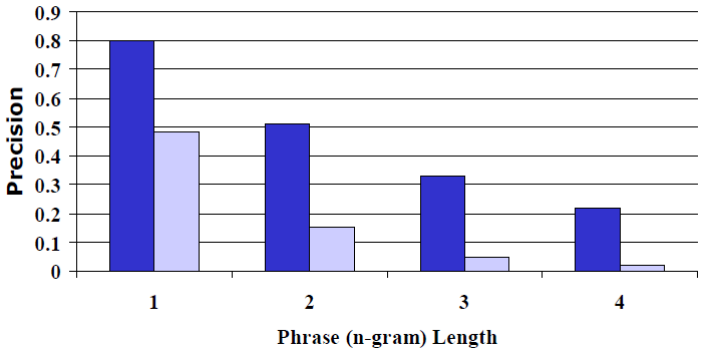
Candidate 1가 Candidate 2에 비해서 확실히 번역이 잘 된 문장이고, Reference의 단어들과 비교해보았을 때, 겹치는 단어가 많다. 이러한 현상은 unigram 뿐만 아니라 n-gram으로 확장했을 때에도, 동일하게 적용된다. 즉, n-gram을 비교할 때, Reference와 겹치는 n-gram이 많을 수록 좋은 MT라는 것이다.

2.1.1 Modified n-gram precision

Candidate: the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

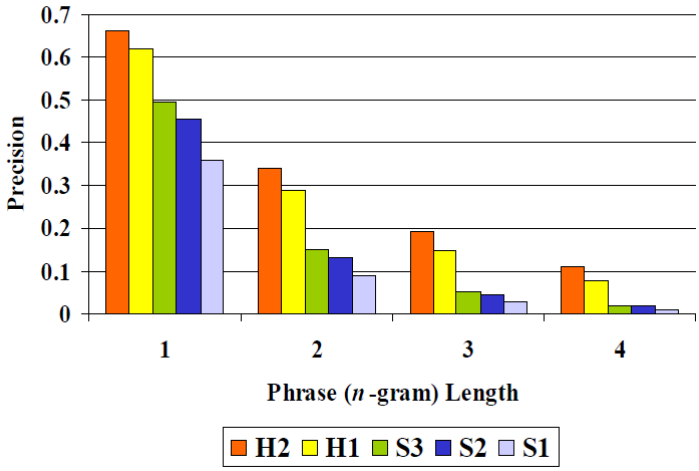
하지만 Candidate과 Reference을 비교하는 naive한 방법으로는, 위와 같이 이상한 MT도 좋다고 판단할 우려가 있다. 따라서 논문에서 modifed unigram precision을 제시한다. Candidate 문장의 단단어가 Reference 문장들과 겹치는 maximum number of times을 계산하고, 이를 Reference 문장 단어의 최댓값으로 나눈다. 위 Candidate 문장에 대해서는, the가 Reference 1 문장과 두 번 겹치고, Reference 2 문장의 단어 개수가 7이기 때문에 modified unigram precision은 2/7이다. 만약 standard nigra precision으로 계산한다면 1이 될 것이다. 같은 논리로 n-gram precision도 계산하면 된다.

Figure 1: Distinguishing Human from Machine



논문에서는 사람이 한 번역과 poor MT를 이용하여 만든 번역에 modified n-gra precision을 이용하여 위와 같이 precision을 계산하였다. 사람이 한 번역이 지속적으로 점수가 훨씬 높음을 확인할 수 있다. 하지만 이러한 결과는 사람간의 좋은 번역도 구분해야 하는 등의 추가 과제가 있다.

Figure 2: Machine and Human Translations



H2가 영어, 중국어에 능통한 사람이고 H1은 서투른 사람이다. 위 결과를 보면, 영어, 중국어에 능통한 사람의 정확도가 그렇지 않은 사람보다 항상 높은 것을 확인할 수 있다.

그림 1, 그림 2를 보면 각 gram 별로 precision을 구했다. 논문에서는 이것들을 하나의 metric으로 합치고자 한다.

2.1.2 Combining the modified n-gram precisions

n-gram의 precision을 합치고자 할 때, gram 별로 정확도가 지수적으로 낮아지는 점을 반영해야 한다. 바로 이 점을 weighted average of te logarithm of modified precisions이 충족한다.
BLEU는 uniform weights의 log 값 평균을 사용하는데, 이는 modified n-gram precisions의 geometric mean을 취하는 것과 동일하다. 논문의 실험에서는, 4-gram 까지의 precisions을 합쳤을 때, 사람의 평가와의 correlation이 가장 컸다고 말한다.

2.2.1 Sentence Length

좋은 MT는 너무 짧지도, 길지도 않은 번역을 해야 한다. n-gram precision이나 modified precision은 이미 이와 유사한 penalization을 하는데, 아래 예시에서는 좋은 Candidate를 탐지하지 못한다.

Candidate: of the
Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the
Party.
Reference 3: It is the practical guide for the army always to heed the directions of the party.

Candidate 문장의 modified unigram precision과 modified bigram precision은 각각 2/2, 1/1이다.

2.2.2 Sentence brevity penalty

Reference 문장보다 더 긴 Candidate 번역은 이미 modified n-gram precision measure에 의해 penalize 된다. 이에 대해서는 더 penalize 할 필요는 없고, 논문 저자는 brevity penalty factor ㅇㄹ 소개한다. 이 brevity penalty는 높은 점수의 Candidate 번역은 Reference 문장과 길이도 맞아야 하고 단어 순서도 맞아야 한다.
brevity penalty가 1.0이라면 Candidate 번역은 Reference 문장과 길이가 동일한 것이다.

2.3 BLEU details

test corpus의 modified precision scores의 geometric mean을 구하고 이 결과에 exponential brevity penalty factor을 곱한다.
먼저 modified n-gram precision의 geometric mean인 p_n을 길이가 N인 gram 까지 구하고 positive weigts인 w_n을 함께 구한다. 다음으로 c를 Candidate 번역의 길이, r을 effective reference corpus length라고 하자. brevity penalty ,BP를 c, r을 이용해서 아래와 같이 구한다.

$$BP = \left\{ \begin{matrix} 1 & \text{if } c > r \\ e^(1 - r/c) & \text{if } c \leq r \end{matrix} \right\} = xy$$

그러면, BLEU는 아래와 같다.

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n log\ p_n \right)$$

BLEU의 ranking behavior은 log을 취한다면 더 분명해 진다.

$$log\ BLEU = min\left(1 - r/c, 0 \right) + \sum_{n=1}^{N} w_n log\ p_n$$

논문 저자들은 baseline으로 $N = 4, w_n = 1/N$을 사용했다고 한다.

3. The BLEU EValuation

BLEU metric 값은 0에서 1의 범위를 가진다. Reference 문장과 똑같이 않는 이상, Candidate 번역이 1의 값을 가지는 경우는 거의 없다. 이러한 이유로 사람의 번역도 1의 점수를 가질 필요는 없다. 또한 중요한 점으로, 문장당 Reference 번역이 많을수록, 점수가 높다는 것이다.

그 이후의 내용은 실험 결과 비교!