BLEU might be Guilty but References are not Innocent

Markus Freitag, David Grangier, Isaac Caswell

Google Research

이은수

Index

목차

Abstract

Abstract

Machine translation(MT)에서 metrics 선택은 고성능 모델에 대해 점점 문제가 커지고 있음

논문에서는 성능 측정 방법과 더불어 references의 품질도 중요하다고 주장함

Paraphrase를 적용해 references의 다양성을 개선하고, 모든 현대적 평가 지표와의 상관관계를 개선함

Introduction

Introduction

Machine translation(MT)의 성능은 최근 몇 년간 크게 향상됐지만, 자동화된 측정 지표의 신뢰성에 의문이 생김

예를 들어 최근 2년간 WMT English→German에 대한 평가는 자동화된 측정 지표와 인간 평가의 순위가 다른 불일치가 관찰됨

그러나 자동화된 평가는 인간 평가에서는 지속 가능하지 않은 속도와 규모를 대신하므로 중요함

Introduction

일반적인 metric는 모델의 출력과 reference 문장간의 일치를 평가해 번역 품질을 측정함

- N-gram 일치 (BLEU)
- 동의어 계산 (METEOR)
- Word representation 고려 (BERTScore)

Reference 문장의 품질은 인간과 자동 평가 사이의 상관관계를 개선하는 데 필수적임

Introduction

번역에 대해

번역자들은 번역어(translationese), 즉 source 문장에 의존하는 번역을 만드는 경향이 있음

그리고 따로 지시하지 않으면 번역자들이 창조적인 번역을 만들지 않음

따라서 metric이 자연적인 번역보다 번역체를 가진 번역에 더 높은 점수를 주도록 편중됨

Introduction

번역체를 피하기 위해 paraphrase reference 문장을 수집해 metric의 문제를 극복하려고 함

1) 동일한 test set에 대해 서로 다른 종류의 reference 문장을 수집하고, 정확도가 높은 모델에서도 인간 측정 기준과 자동 평가 사이에 강한 상관관계를 가질 수 있음
2) Paraphrase reference 문장을 수집해 자동 평가에서 reference로 사용될 때 다양성, 정확성, 자연성, 인간 판단과의 상관성 측면에서 여러 흥미로운 성질을 가짐
3) Multi-reference BLEU보다 더 효과적인 대안을 제시함
4) 다른 종류의 번역을 만드는 연구를 장려하기 위해 다양한 reference 문장을 제시함

Introduction

Paraphrase란?

원 문장의 뜻을 변형하지 않으면서 다른 말이나 문장으로 새롭게 바꾸어 표현한 것

1. 동의어(synonym) / 의미(definition) 활용
2. 파생어(word derivative) 활용
3. 수동태 및 능동태 형태 전환
4. 문장 병합

등이 있다!

"[영어에세이] Paraphrasing이 뭔가요? Paraphrase 하는 방법/연습/예시 및 훈련", Sourbear, https://m.blog.naver.com/sourbear/221848710320

Related Work

Related Work

Metric은 일반적으로 output과 reference간의 중첩을 측정함

매우 다양한 metric이 제안되었고, 그 중 BLEU는 가장 일반적인 측정 기준

1) **BLEU**는 N-gram에 대한 기하학적 평균을 측정하고, 짧은 번역을 방해하는 penalty를 제공함
2) **NIST**는 BLEU와 유사하지만 가중치가 상승하는 희귀한 N-gram을 고려함
3) **TER**은 reference와 비교할 때 수정을 위한 거리를 측정함
4) **METEOR**는 동의어를 고려해 정확한 match 이상의 N-gram 보상을 제안함
5) **BERTScore**는 문맥화된 word embedding을 사용할 것을 제안함

Related Work

이 연구에서는...

언어 모델에 기반해 metric을 측정하는 최근의 관련 연구와 달리, 이와 독립적으로 평가가 이루어지기를 바람

대신 번역자로부터 멀어지기 위해 paraphrase를 사용해 더 다양한 reference를 수집하는 것을 목적으로 함

Collecting High Quality and Diverse References

Collecting High Quality and Diverse References

Increasing reference quality

추가로 reference 문장을 얻기 위해 두 가지 접근법을 시도함

1) 전문 번역 서비스에 추가 번역 제공을 요청함
2) 다른 언어학자들에게 질문하면서 기존의 references를 paraphrase하기 위해 같은 서비스를 이용함

CAT 도구(사전, 번역 메모리, MT) 등을 불허하고, source 문장을 복사할 수 없도록 함

Collecting High Quality and Diverse References

Diversified, natural references through paraphrasing

100문장 샘플에 대한 초기 실험에서 언어학자들에게 문장을 paraphrase해줄 것을 요청했지만, 아주 사소한 변경만이 있었음

Paraphrase 시 동의어 및 다른 문장 구조를 사용할 것을 제안해 paraphrase된 문장이 서로 아주 다른 구조를 가지게 됨

"가능한 한 많이" paraphrase하는 것은 때로는 적절한 정도를 넘어서 과도할 수 있기 때문에, 생성된 paraphrase 문장에 대해 등급을 매김

Task: Paraphrase the sentence as much as possible:

To paraphrase a source, you have to rewrite a sentence without changing the meaning of the original sentence.

1. Read the sentence several times to fully understand the meaning
2. Note down key concepts
3. Write your version of the text without looking at the original
4. Compare your paraphrased text with the original and make minor adjustments to phrases that remain too similar

Please try to change as much as you can without changing the meaning of the original sentence.

Some suggestions:

1. Start your first sentence at a different point from that of the original source (if possible)
2. Use as many synonyms as possible
3. Change the sentence structure (if possible)

Figure 1: Instructions used to paraphrase an existing translation *as much as possible*.

Paraphrase 작업 지침서

Source	The Bells of St. Martin's Fall Silent as Churches in Harlem Struggle .
Translation	Die Glocken von St. Martin verstummen , da Kirchen in Harlem Probleme haben .
Paraphrase	Die Probleme in Harlems Kirchen lassen die Glocken von St. Martin verstummen .
Paraphrase	Die Kirchen in Harlem kämpfen mit Problemen , und so läuten die Glocken von St. Martin nicht mehr .

Table 1: Reference examples of a typical translation and two different paraphrases of this translation. The paraphrases are not only very different from the source sentence (e.g. sentence structure), but also differ a lot when compared to each other.

Paraphrase 문장 예시

Experiments

Experiments

Correlation with Human Judgement

BLEU의 순위상관관계 테이블

- **WMT**: 번역 데이터 세트
- **AR**: WMT 데이터 세트에 대한 추가 reference
- **WMT.p**: WMT 데이터 세트에 대한 paraphrase
- **AR.p**: AR 데이터 세트에 대한 paraphrase
- **HQ**: 정확도가 높은 문장만 선택한 데이터 세트

AR, WMT.p, AR.p는 모두 원래 데이터 세트인 WMT보다 높은 상관관계를 기록한다

Full Set (22)	Reference	ρ	τ
single ref	WMT	0.88	0.72
	AR	0.89	0.76
	WMT.p	**0.91**	**0.79**
	AR.p	0.89	0.77
single ref	HQ(R)	**0.91**	0.78
	HQ(P)	**0.91**	0.78
	HQ(all 4)	**0.91**	**0.79**
multi ref	AR+WMT	0.90	0.75
	AR.p+WMT.p	0.90	**0.79**
	all 4	0.90	0.75

Table 3: Spearman's ρ and Kendall's τ for the WMT2019 English→German official submissions with human ratings conducted by the WMT organizers.

Experiments

단, 상관관계 점수가 이미 상대적으로 높게 나타난다는 점을 유의해야 함

따라서 작은 숫자의 증가는 훨씬 더 큰 순위 향상에 해당할 수 있음

top-k로 본다면...

1) WMT 2019 데이터 세트는 human rating과는 상관관계가 낮음
2) Paraphrase된 데이터 세트에 대해서는 그렇지 않은 것과 비교해 모두 높은 점수를 기록함

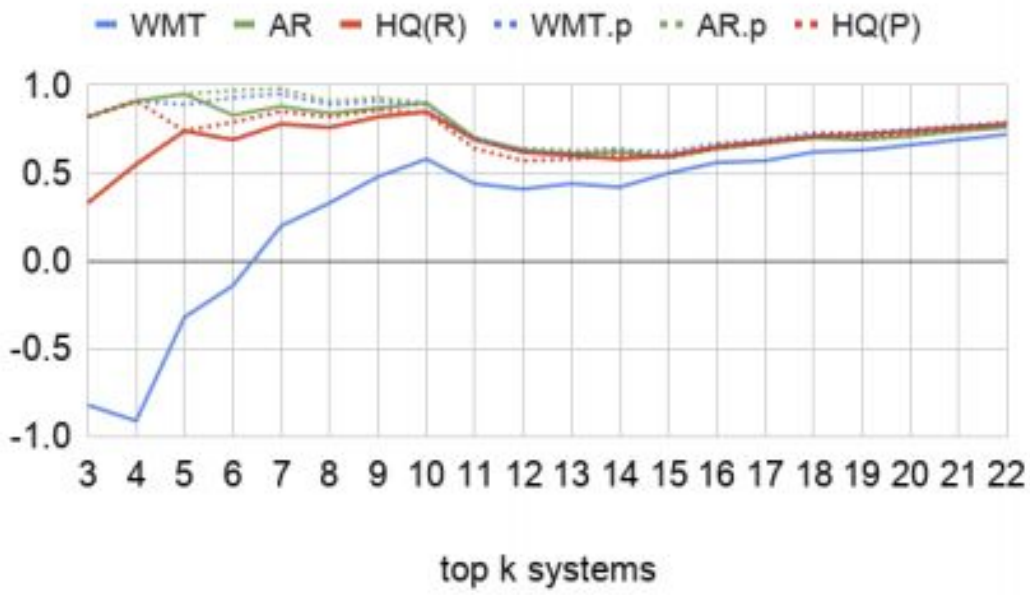

Figure 2: Kendall's τ correlation of BLEU for the best k systems (based on human ratings).

Experiments

Alternative Metrics

Paraphrase된 데이터 세트는 여러 metrics에서 기존 데이터 세트 (WMT)보다 높은 점수를 기록함

특히 word embedding을 사용하는 Yisi-1은 reference의 더 높은 정확도로 큰 이득을 얻는 것으로 보임

metric	WMT	HQ(R)	WMT.p	HQ(P)	HQ(all)
BLEU	0.72	0.78	**0.79**	**0.79**	**0.79**
1 - TER	0.71	**0.74**	0.71	0.67	**0.74**
chrF	0.74	0.81	0.78	**0.82**	0.78
MET	0.74	**0.81**	**0.81**	**0.81**	0.80
BERTS	0.78	**0.82**	**0.82**	**0.82**	0.81
Yisi-1	0.78	0.84	0.84	**0.86**	0.84

Table 5: WMT 2019 English→German: Correlations (Kendall's τ) of alternative metrics: BLEU, 1.0 - TER, chrF, METEOR, BERTScore, and Yisi-1.

Measuring Translationese

Measuring Translationese

번역어는 원본 언어의 전형적인 특징을 유지함

따라서 번역에 존재하는 번역자의 정도를 수량화하려고 시도하는 metric이 제안된 바 있음

어휘적 다양성(lexical variety)과 **어휘적 밀도(lexical density)**라는 두 가지 지표로 어휘적 단순성을 정량화하고, source 문장의 간섭을 측정하기 위해 **길이 다양성(length variety)**을 계산함

Measuring Translationese

Lexical Variety

더 적은 수의 고유 토큰/단어를 사용할 때 더 많은 번역어가 사용된다

$$lex\_variety = \frac{number\ of\ types}{number\ of\ tokens}$$

Measuring Translationese

Lexical Density

번역어가 원문에 비해 어휘적으로 단순하고 content words(부사, 형용사, 명사, 동사)의 비율이 낮음

$$lex\_density = \frac{number\ of\ content\ words}{number\ of\ total\ words}$$

Measuring Translationese

Length Variety

Machine translation과 인간 모두 source 문장의 구조조정을 피하고 문장 구조를 고수하는 경향이 있음

이것은 source 문장과 비슷한 길이의 번역을 초래함

따라서 source-target 쌍 (x, y)에 대해 문장 길이의 절대값을 측정함

$$len\_variety = \frac{||x| - |y||}{|x|}$$

Measuring Translationese

모든 지표에서 de.tr은 많은 번역체 문장을 출력한다는 것을 확인하면서 가장 낮은 점수를 받음

반면 paraphrase는 어휘적 다양성은 낮지만 번역된 문장과 원래 독일어 문장에 비해 어휘적 밀도와 길이적 다양성이 훨씬 높음

가능한 한 paraphrase를 함으로써 많은 번역체를 제거할 수 있었다는 것을 보여줌

	Lex. Var.	Lex. Density	Len. Var.
de.orig	**0.534**	**0.398**	**0.134**
de.tr	0.509 *-4.6%*	0.391 *-1.8%*	0.131 *-2.2%*
de.orig.p	0.513 *-3.9%*	0.408 ***+2.0%***	0.195 ***+45%***
de.tr1.p	0.522 *-2.2%*	0.400 ***+0.5%***	0.196 ***+46%***

Table 10: Measuring the degree of translationese, reporting percent difference wrt. to de.orig. Higher lexical variety, lexical density, and length variety imply less translationese sentences. Values at or exceeding those of natural text are bolded.

Conclusions

Conclusions

Machine translation의 자동화된 평가의 신뢰성에 대한 reference의 품질이 미치는 영향에 관한 연구를 제시함

Q & A

End of presentation