第11回最先端NLP勉強会（2019.9.27）

Beyond BLEU: Training Neural Machine Translation with Semantic Similarity

John Wieting[1], Taylor Berg-Kirkpatrick[2], Kevin Gimpel[3], and Graham Neubig[1]

論文: https://www.aclweb.org/anthology/P19-1427

発表動画: http://www.livecongress.it/aol/indexSA.php?id=8FD8C680

紹介者: 笹野遼平（名大）

Summary

- ニューラル機械翻訳(NMT)の最適化で用いるBLEUに代わる報酬関数として意味的類似度(semantic similarity)に基づく指標SIMILEを提案

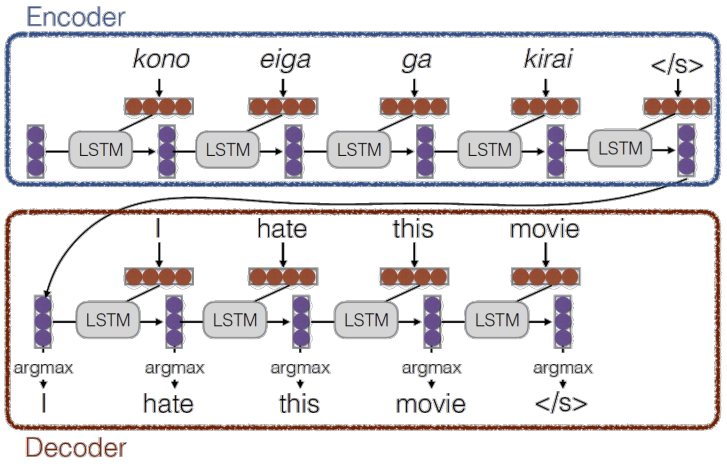
$$\text{SIMILE} = \text{LP}(r,h)^{\alpha}\text{SIM}(r,h)$$

Length penalty　　Semantic Similarity

- BLEU＆人手評価の結果が改善
- 連続的な報酬関数であるため収束も高速化

Minimum Risk Training for Neural Machine Translation [Shen+'16]

- NMTモデルを評価指標(e.g., BLEU)に対して直接最適化する枠組み



出力候補**u**の参照翻訳**t**に対するコスト(e.g., 1-BLEU)

$$\mathcal{L}_{\text{Risk}} = \sum_{\mathbf{u} \in \mathcal{U}(\mathbf{x})} \text{cost}(\mathbf{t}, \mathbf{u}) \frac{p(\mathbf{u}|\mathbf{x})}{\sum_{\mathbf{u}' \in \mathcal{U}(\mathbf{x})} p(\mathbf{u}'|\mathbf{x})}$$

出力候補の集合

出力候補で正規化したuの翻訳確率

BLEU score

- 機械翻訳の代表的な評価尺度
  - (評価コーパス全体で)翻訳文と参照訳を比較したときの翻訳文のn-gramの一致率(p_n)に基づく(=precisionベース)
  - 短い出力にペナルティを与えることで短い翻訳を抑制

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

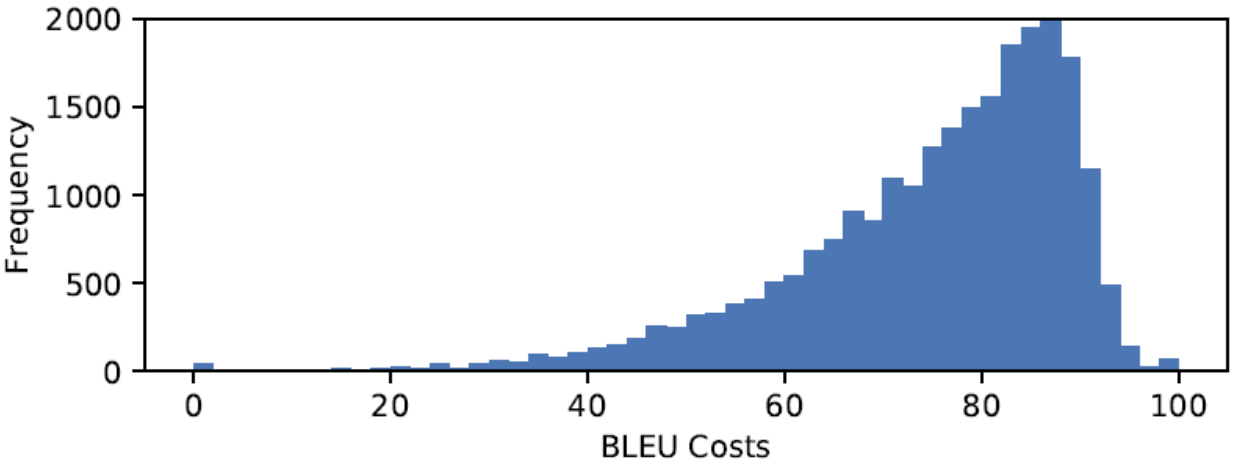
Brevity Penalty: $$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

- sentence-level BLEU: 文単位で算出したBLEU
  - 本論文では基本的にはこちらを利用

BLEUの問題点

1. 参照要約と異なる語彙が使用された場合、
   意味的に正しい出力にもペナルティを与える

2. 出力値が離散的＆分布の偏りが大きい
   ⇒ 訓練時に利用できる情報量が限定的



BLEUで上手く評価できない例

- ほぼ同じ意味であるものの使用語彙が異なる場合
  - Ref: I don't know how to explain – it's really unique
  - S1 : I *do not* know how to explain *it* – *it is* really unique. s-BLEU:39.1
  - S2 : I don't know how to explain – *it is* really unique. s-BLEU:78.3

- 類似文字列であるものの意味が大きく異なる場合
  - Ref: All that stuff does take a toll
  - S1 : *None* of this takes a toll. s-BLEU:26.0
  - S2 : All of this is certain to take its toll. s-BLEU:18.9

⇨ 意味的類似度に基づく報酬(コスト)関数を購入

文類似度尺度: SIM [Wieting&Gimpel'08]

- 300次元のsubword embeddingの平均ベクトル$g(\cdot)$の余弦類似度で算出（語順等は考慮していない←LSTMが頑張る？）

- 学習時は言い換えペア$<s, s'>$に対し下記を最小化

$$\ell(s, s') = \max(0, \delta - \cos(g(s), g(s')) + \cos(g(s), g(t)))$$

  - t: 負例（mini-batch中でもっとも類似度の大きいもの）
  - δ: 正例と負例の間のマージン

SIMILE = SIM + LP(Length Penalty)

$$\text{SIMILE} = \text{LP}(r,h)^{\alpha}\text{SIM}(r,h)$$

- SIMは長い文、特に繰り返し(repetition)を含む文の出力を抑制できない
- LPを導入(BLEUにおけるBPに類似)

$$\text{LP}(r,h) = e^{1-\frac{\max(|r|,|h|)}{\min(|r|,|h|)}} \qquad \text{cf. } \text{BP}(r,h) = e^{1-\frac{|r|}{|h|}}$$

  - 長さが"異なる"ことに対しペナルティ

実験

- 4言語対で実験
  - {de, cs, ru, tr} ⇔ en
- 評価尺度
  - corpus-level BLEU
  - SIM (SIMILEでないことに注意⇒意味内容の一致度合いに着目)
  - 人手評価(200文を0-5の6段階で評価)
- 比較手法
  - MLE: Maximum likelihood with label smoothing
  - BLEU: Minimum risk training with $1 - \text{BLEU}$
  - SIMILE: Minimum risk training with $1 - \text{SIMILE}$
  - Half: Minimum risk training with $1 - \frac{1}{2}(\text{BLEU} + \text{SIMILE})$

Lang.	Train	Valid	Test
cs-en	218,384	6,004	2,983
de-en	284,286	7,147	2,998
ru-en	235,159	7,231	3,000
tr-en	207,678	7,008	3,000

実験結果

- BLEUとSIMによる評価

	de-en		cs-en		ru-en		tr-en	
Model	BLEU	SIM	BLEU	SIM	BLEU	SIM	BLEU	SIM
MLE	27.52	74.96	17.02	67.18	17.92	70.24	14.47	63.52
BLEU	27.95‡	86.93‡	17.29‡	81.92‡	17.92	84.63‡	15.00‡	80.30‡
SIMILE	**28.28**†‡	**87.32**†‡	**17.51**†‡	**82.12**†‡	18.23†‡	**85.12**†‡	15.28†‡	**81.04**†‡
Half	28.24†‡	87.11†‡	17.50†‡	82.11†‡	**18.24**†‡	85.10†‡	**15.34**†‡	80.64†‡

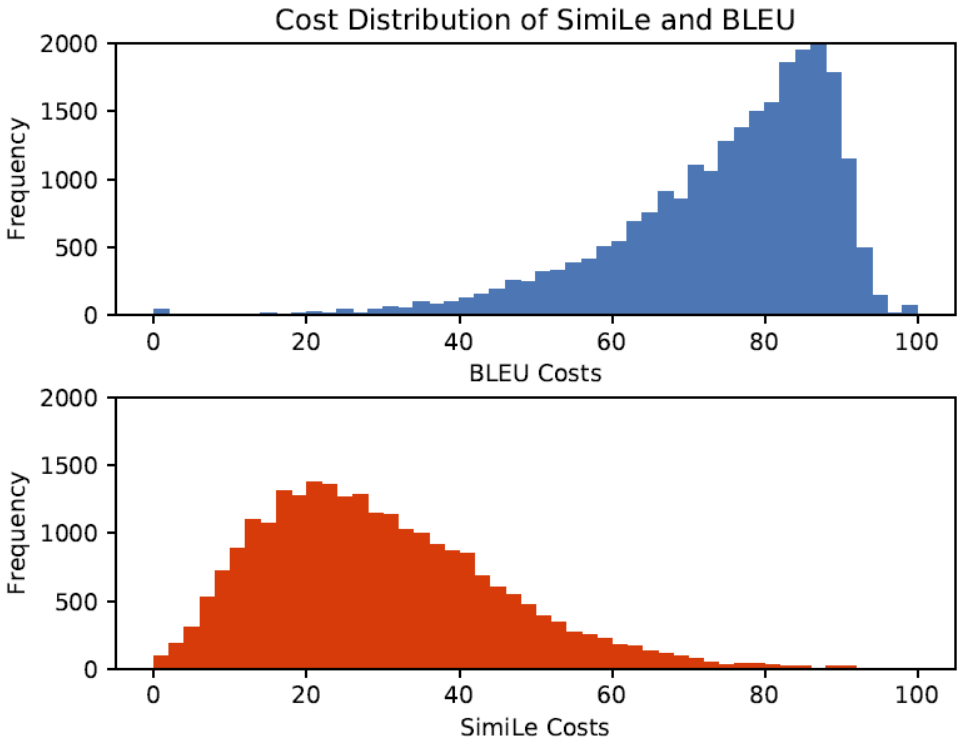
  - SIMILEはいずれの評価尺度に対しても最も良いスコア
    - もともとの目的はBLEUの改善ではなく人手評価の改善（welcome surprise）
  - BLEUもMLEと比べSIMが改善

- 人手評価(0-5の6段階)
  - トルコ語以外はSIMILEが最高精度
  - トルコ語が最もBLEUのスコアが低いことから事前にある程度の翻訳精度が必要かもと分析（やや疑問）

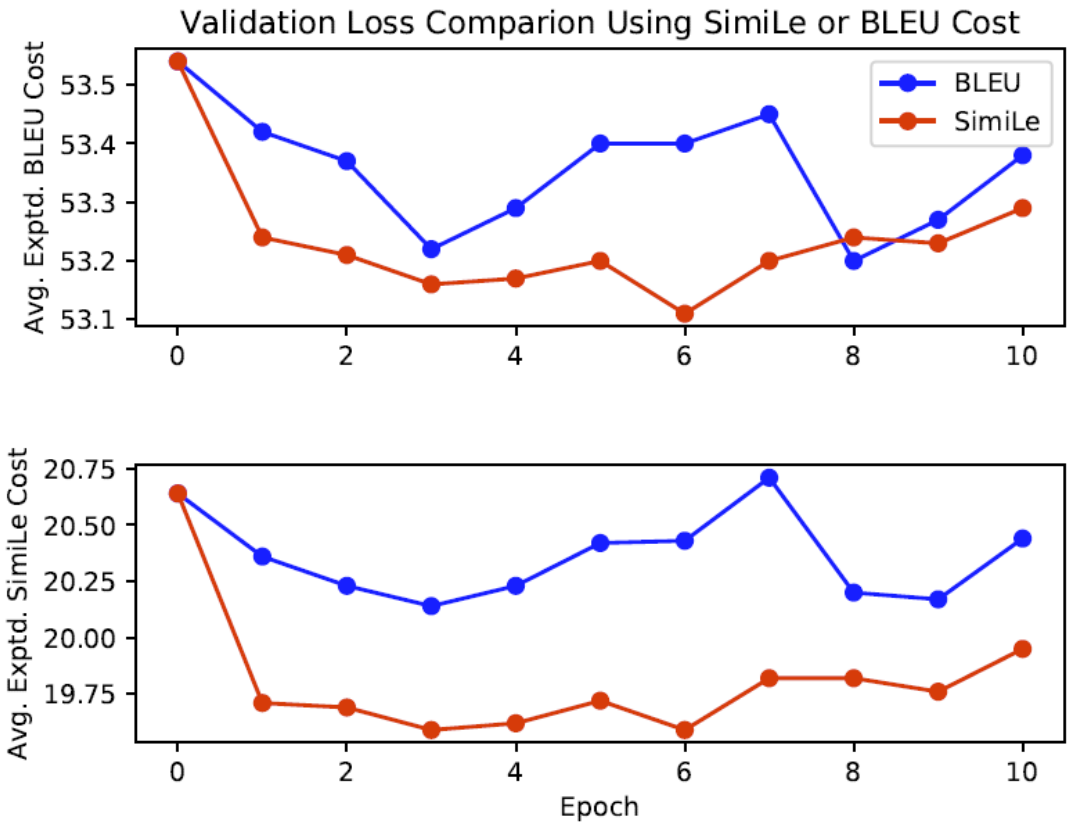
	Avg. Score		
Lang.	MLE	BLEU	SIMILE
cs-en	0.98	0.90	**1.02**†
de-en	0.93	0.85	**1.00**†
ru-en	1.22	1.21	**1.31**†‡
tr-en	0.98*	**1.03***	0.78

Quantative Analysis (de-en)

- BLEUとSIMILEに基づくコストの分布を比較するとBLEUの方が偏りが大きい
- ビーム幅8の用例間28ペア間で、スコアが異なる割合/平均スコア差
  - BLEU: 85.1% / 4.3
  - SIMILE: 99.0% / 4.8
- 訓練時により多くの情報を利用できている



Validation Loss



- SIMILEを用いた場合のほうが収束が早く、かつ、より小さなValidation Lossに到達

$\Delta\mathrm{BLEU}\ (= \mathrm{BLEU}_{\mathrm{BLEU}} - \mathrm{BLEU}_{\mathrm{SIM}})$と $\Delta\mathrm{SIM}\ (= \mathrm{SIM}_{\mathrm{BLEU}} - \mathrm{SIM}_{\mathrm{SIM}})$の差が大きい例

Reference	Workers are beginning to clean up workers .
BLEU system	Workers have begun to clean up in Rszke.
SIM system	In Rszke, workers are beginning to clean up.
ΔBLEU	3.2
ΔSIM	-26.3
Reference	All that stuff sure does take a toll.
BLEU system	None of this takes a toll .
SIM system	All of this is certain to take its toll .
ΔBLEU	7.1
ΔSIM	-22.7
Reference	Another advantage is that they have fewer enemies.
BLEU system	Another benefit: they have less enemies.
SIM system	Another advantage: they have fewer enemies.
ΔBLEU	-33.8
ΔSIM	-9.6
Reference	I don't know how to explain - it's really unique.
BLEU system	I do not know how to explain it - it is really unique.
SIM system	I don't know how to explain - it is really unique.
ΔBLEU	-39.1
ΔSIM	-2.1

逆？

Summary

- ニューラル機械翻訳(NMT)の最適化で用いるBLEUに代わる報酬関数として意味的類似度(semantic similarity)に基づく指標SIMILEを提案

$$\text{SIMILE} = \text{LP}(r,h)^{\alpha}\text{SIM}(r,h)$$

Length penalty　Semantic Similarity

- BLEU＆人手評価の結果が改善
- 連続的な報酬関数であるため収束も高速化

雑感

- BLEUによる評価でも精度向上してるのは凄い
  - “welcome surprise”と表現されており想定外な結果？
  - 意味が類似するとスコアが高くなるのは直感的に良さそう
  - 報酬関数が連続的（BLEUは離散的）なのもポイント？

- 人手評価を見るとMLEより良いかやや疑問
  - ４言語対中３言語対で改善してるものの平均値は同じ
  - そもそも報酬関数としてBLEUを用いるのは良いのか？[Wu+'16]

	Avg. Score		
Lang.	MLE	BLEU	SIMILE
cs-en	0.98	0.90	**1.02**†
de-en	0.93	0.85	**1.00**†
ru-en	1.22	1.21	**1.31**†‡
tr-en	0.98*	**1.03***	0.78
Avg.	1.03	1.00	1.03