An Awkward Disparity between BLEU / RIBES Scores and Human Judgements in Machine Translation

Liling Tan, Jon Dehdari¹ and Josef van Genabith¹
Universität des Saarlandes
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI)¹
liling.tan@uni-saarland.de, {first.last_name}@dfki.de

Introduction

• MT metrics criticized for various reasons
  (Babych and Hartley, 2004; Smith et al. 2014; Graham et al. 2015)

Hypothesis 1:
Appeared calm when he was taken to the American plane, which will to Miami, Florida.

Hypothesis 2:
which will he was, when taken appeared calm to the American plane to Miami, Florida

Reference:
Orejuela appeared calm as he was led to the American plane which will take to Miami, Florida.

• Low BLEU != Bad MT (Callison-Burch et al. 2006)
• Higher BLEU -> Better MT (c.f. WMT, WAT, IWSLT, OpenMT)

BLEU: (Papineni et al. 2002)
• Precision based
• Weak recall penalty
• Disregards order

Source:
여리한을병합하기 위해서는, 다음과 0.009% 이상할 수는 의미를 범위합니다.

Hypothesis:
このような 문제를 해결하기 위해서는, 앞의 0.005%의 경우에 문제가 발생하지 않습니다.

Baseline:
このような 문제를 해결하기 위해서는, 앞의 0.005%의 경우에 문제가 발생하지 않습니다.

Reference:
여리한을병합하기 위해서는, 다음과 0.009% 이상할 수는 의미를 범위합니다.

RIBES (Isozaki et al. 2014)
• Kendall Tau prior on unigram
• Overcomes reordering
• Adequacy not measured
• Correlates with BLEU (naturally)

Hypothesis
RIBES: 94.04
BLEU: 53.3
HUMAN: -5

Baseline
RIBES: 86.33
BLEU: 58.8
HUMAN: 0

System Setup + Results

Organizers:
RIBES = 94.13 ; BLEU = 69.22 ; HUMAN = 0.0

Ours:
RIBES = 95.15 ; BLEU = 85.23 ; HUMAN = -17.75 !!!

Note: This is our Unicode2String submission for KO->JA patent subtask in WAT 2015; the other results of other subtasks are presented in Tan and Bond (2014) and Tan et al. (2015).

Conclusion

• Higher BLEU/RIBES correlates with +ve HUMAN, not -ve HUMAN
• Minor lexical diff. cause huge diff. in BLEU, RIBES mostly measures fluency
• Minor metric score diff. not reflecting major translation inadequacy
• Higher BLEU = Better MT

References
• Bogdan Babych and Anthony Hartley. Extending the BLEU MT evaluation method with frequency weightings. In ACL.
• Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. Findings of the workshop on statistical machine translation. In WMT.
• Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of Bleu in machine translation research. In EACL.
• Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. Report on the IWSLT evaluation campaign, IWSLT. In IWSLT.
• Yvette Graham, Timothy Baldwin, and Nitika Mathur. Accurate evaluation of segment-level machine translation metrics. In ACL.
• Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. Automatic evaluation of translation quality for distant language pairs. In EMNLP.
• Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. Overview of the workshop on Asian translation. In WAT.
• Liling Tan and Francis Bond. Manipulating in-put data in machine translation. In WAT.
• Liling Tan, Josef van Genabith, and Francis Bond. Passive and pervasive use of bilingual dictionary in statistical machine translation. In HyTra.