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

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

UNIVERSITÄT DES SAARLANDES

Marie Curie Actions

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

Same BLEU?!

Reference:
Orejuela appeared calm as he was led to the American plane which will take him 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
· Crude ngram weights
· Over-sensitive to minor difference

Source:
이러한작용을발휘하기위해서는, 각각 0.005% 이상함유하는것이바람직하다.

Hypothesis:
このような作用を発揮するためには、夫々０．００５％以上含有することが好ましい。

Baseline:
このような作用を発揮するためには、それぞれ０．００５％以上含有することが好ましい。

Reference:
このような作用を発揮させるためには、夫々０．００５％以上含有させることが好ましい。

Hypothesis	Baseline
P_1 : 90.0	P_1 : 84.2
P_2 : 78.9	P_2 : 66.7
P_3 : 66.7	P_3 : 47.1
P_4 : 52.9	P_4 : 25.0
BP: 0.905	**BP:** 0.854
BLEU: 64.03	**BLEU:** 43.29
HUMAN: -5	**HUMAN:** 0

Acknowledgements: The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n ° 317471.

RIBES (Isozaki et al. 2014)

· Kendall Tau prior on unigram
· Overcomes reordering
· Adequacy not measured
· Correlates with BLEU (*naturally*)

Source:
T용융(DSC) = 89.9℃; T결정화(DSC) = 72℃(5℃/분에서DSC 로측정).

Hypothesis:
Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89．９結晶化度（Ｔ）。

Baseline:
Ｔ溶融（ＤＳＣ）＝８９．９℃；Ｔ結晶化（ＤＳＣ）＝７２℃（５℃／分でＤＳＣで測定）。

Reference:
Ｔｍｅｌｔ（ＤＳＣ）＝８９．９℃；Ｔｃｒｙｓｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣを用いて測定）。

Hypothesis
RIBES: 94.04
BLEU: 53.3
HUMAN: -5

Baseline
RIBES: 86.33
BLEU: 58.8
HUMAN: 0

System Setup + Results

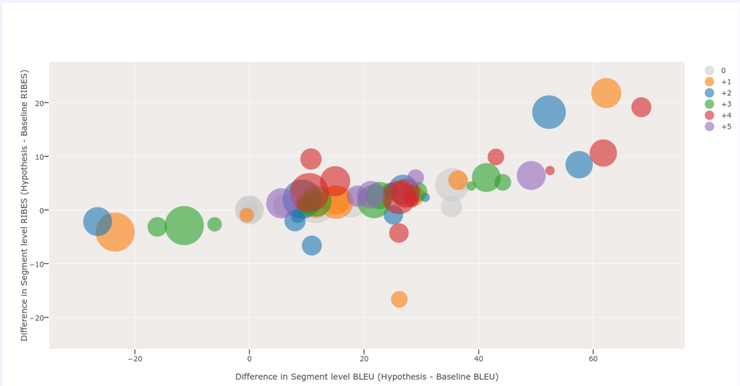
Parameters	Organizers	Ours
Input document length	40	80
Korean tokenizer	MeCab	KoNLPy
Japanese tokenizer	Juman	MeCab
LM *n*-gram order	5	5
Distortion limit	0	20
Quantized & binarized LM	no	yes
`devtest.txt` in LM	no	yes
Binarized phrase tables	no	yes
MERT runs	1	2

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

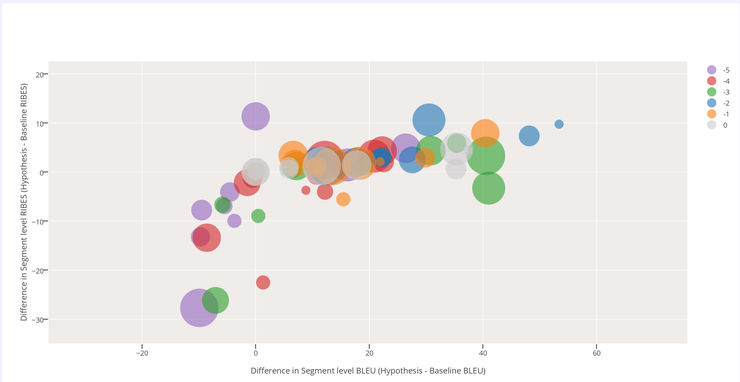
· **Ours** :
RIBES = 95.15 ; BLEU = **85.23** ; HUMAN = **-17.75 !!!**

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)

Meta-Evaluation



Bubble Graph of Diff. Hypothesis - Baseline BLEU against RIBES for ***Positive*** HUMAN Scores



Bubble Graph of Diff. Hypothesis - Baseline BLEU against RIBES for ***Negative*** HUMAN Scores

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. 2004. Ex- tending 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. 2015. Findings of the 2015 workshop on statistical machine translation. In WMT.
· Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of Bleu in machine translation research. In EACL.
· Mauro Cettolo, Jan Niehues, Sebastian StÃijker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In IWSLT.
· Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In ACL.
· Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In EMNLP.
· Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. 2015. Overview of the 2nd workshop on Asian translation. In WAT.
· Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In ACL.
· Liling Tan and Francis Bond. 2014. Manipulating in- put data in machine translation. In WAT.
· Liling Tan, Josef van Genabith, and Francis Bond. 2015. Passive and pervasive use of bilingual dictio- nary in statistical machine translation. In HyTra.