An Awkward Disparity between BLEU / RIBES and Human Judgment in MT

Liling Tan, Jon Dehdari and Josef van Genebith
Saarland University, Germany

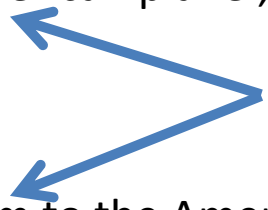
@alvations

Introduction

- **There's always *a bone to pick* on MT evaluation metrics** (Babych and Hartley, 2004; Callison-Burch et al. 2006; 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 .



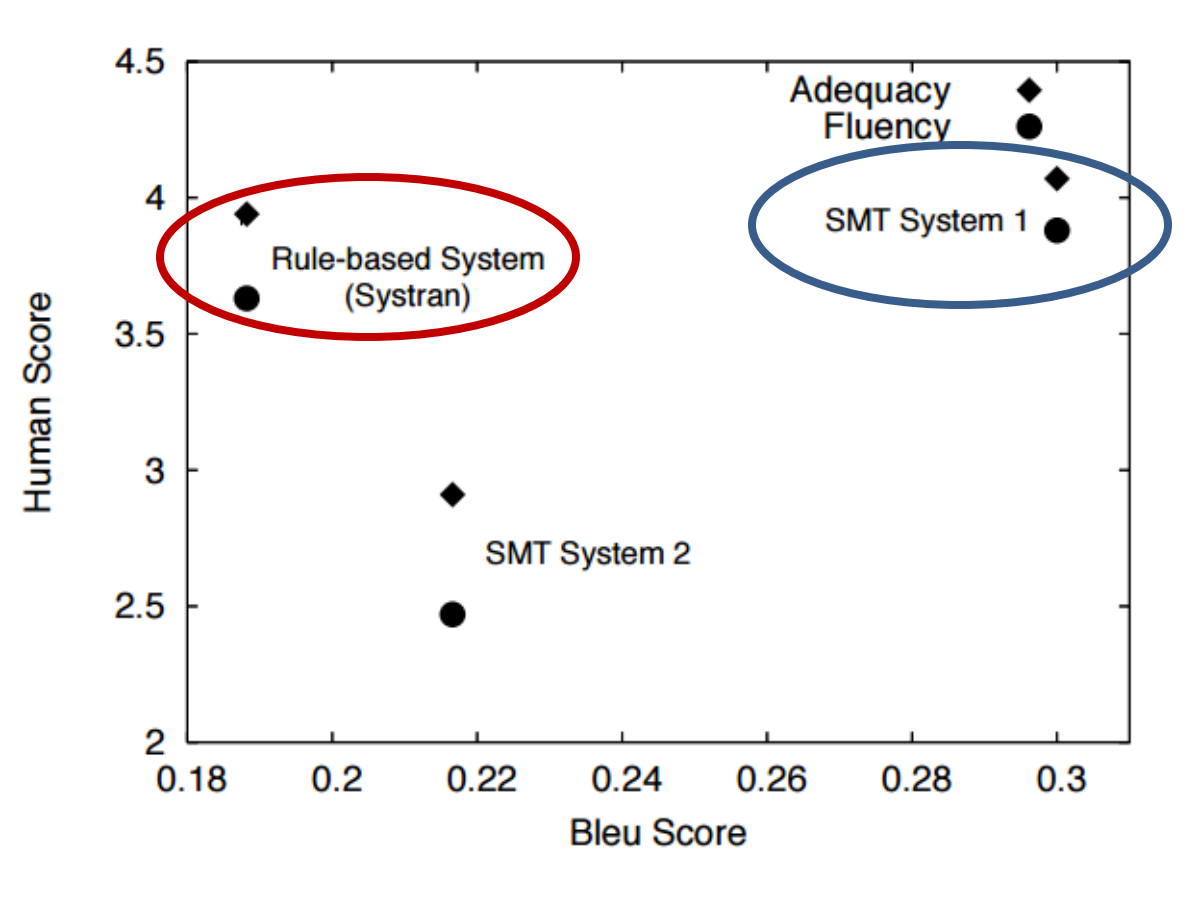
Reference:
Orejuela appeared calm as he was led to the American plane which will take him to Miami , Florida .

Introduction

- ***"Conventional"* wisdom**:
  - ***lower BLEU not necessarily worse translation***
    (Callison-Burch et al. 2006)
  - ***higher BLEU = better translation***

    (Callison-Burch et al. 2006; Nakazawa et al., 2014;
    Cettolo et al., 2014; Bojar et al., 2015)

Introduction



Callison-Burch et al. (2006) meta-evaluation on 2005 NIST MT Eval

Introduction

- ***"Conventional"* wisdom**:
  - ***lower BLEU not necessarily worse translation***
    (Callison-Burch et al. 2006)
  - ***higher BLEU = better translation***

    (Callison-Burch et al. 2006; Nakazawa et al. 2014;
    Cettolo et al. 2014; Bojar et al. 2015)

But is higher BLEU = better translation true?

BLEU

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

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

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

Penalize if the length of the hypothesis is too long

Count the proportion of *n*-grams that appears in hypothesis and reference

BLEU (in practice)

$$\text{BLEU} = \text{BP} \times (p_1 p_2 p_3 p_4)^{1/4}$$

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

Penalize if the length of the hypothesis is too long

$$p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}$$

Count the proportion of *n*-grams that appears in hypothesis and reference

BLEU

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**

RIBES

Source:
T용 융(DSC) = 89.9℃; T결 정 화(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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)

Hypothesis	Baseline
RIBES: 94.04	RIBES: 86.33
BLEU: 53.3	BLEU: 58.8
HUMAN: -5	HUMAN: 0

System Level HUMAN

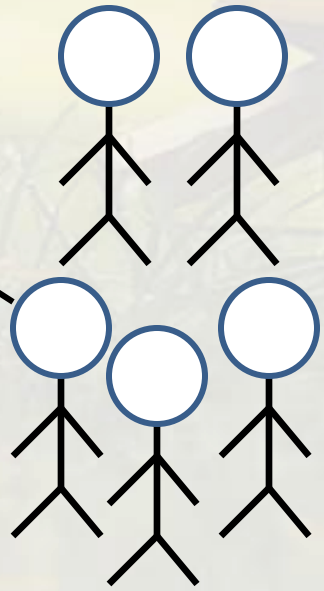
Source:
T용 융(DSC) = 89.9℃; T결 정 화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)



Hyp < Base

= 0 < 5

= -1 HUMAN

System Level HUMAN

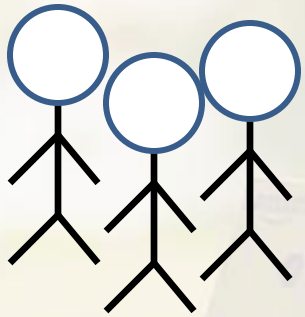
Source:
T용 융(DSC) = 89.9℃; T결 정 화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

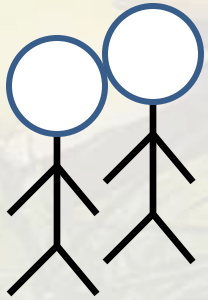
Hypothesis:
Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89．９結晶化度（Ｔ）。

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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)





Hyp > Base

= 3 > 2

= +1 HUMAN

System Level HUMAN

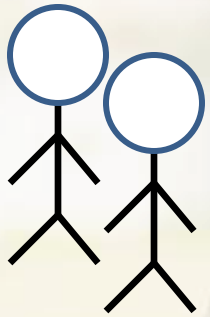
Source:
T용 융(DSC) = 89.9℃; T결 정 화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

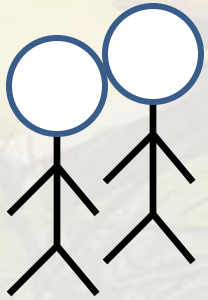
Hypothesis:
Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89．９結晶化度（Ｔ）。

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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)





Hyp == Base
= +0 HUMAN

Segment Level HUMAN

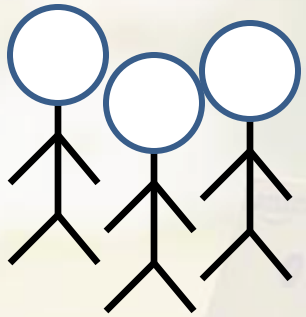
Source:
T용 융(DSC) = 89.9℃; T결 정 화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

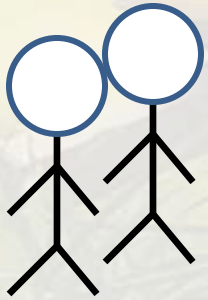
Hypothesis:
Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89．９結晶化度（Ｔ）。

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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)





#Hyp - #Base

= 3 -2

= +1 HUMAN

Segment Level HUMAN

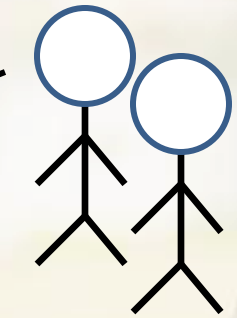
Source:
T용 융(DSC) = 89.9℃; T결 정 화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

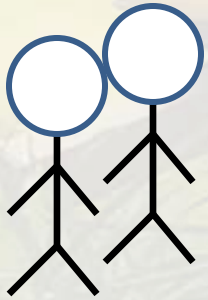
Hypothesis:
Ｔｍｅｌｔ（ＤＳＣ）＝７２℃（５℃／分でＤＳＣ測定（ＤＳＣ）＝89．９結晶化度（Ｔ）。

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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)





#Hyp - #Base
= 2 -2
= 0

Segment Level HUMAN

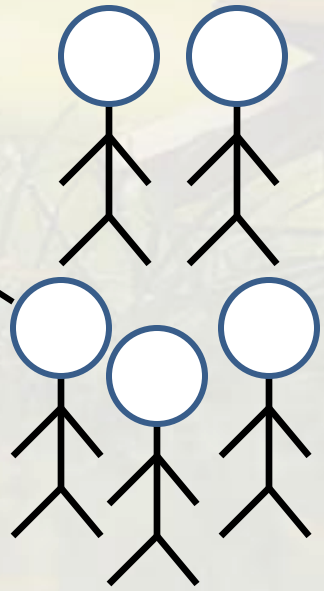
Source:
T용 융(DSC) = 89.9℃; T결 정 화(DSC) = 72℃(5℃/ 분에서DSC 로측정) .

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）。

Source/Reference English Gloss:
Tmelt (DSC) = 8 9. 9 °C; Tcryst (DSC) = 7 °C (measured using DSC at 5 °C / min)



#Hyp - #Base

= 0 - 5

= -5 HUMAN

Experiment Setup

(Our WAT Submission)

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

Results

(Our WAT Submission)

Systems	RIBES	BLEU	HUMAN
Organizers' PBMT baseline	94.13	69.22	0.0
Our replica baseline	94.29	70.23	**+3.50**
Ours (MERT 1)	95.03	84.26	-
Ours (MERT 2)	**95.15**	**85.23**	-17.75

+15 BLEU -> -17.75 HUMAN !!!

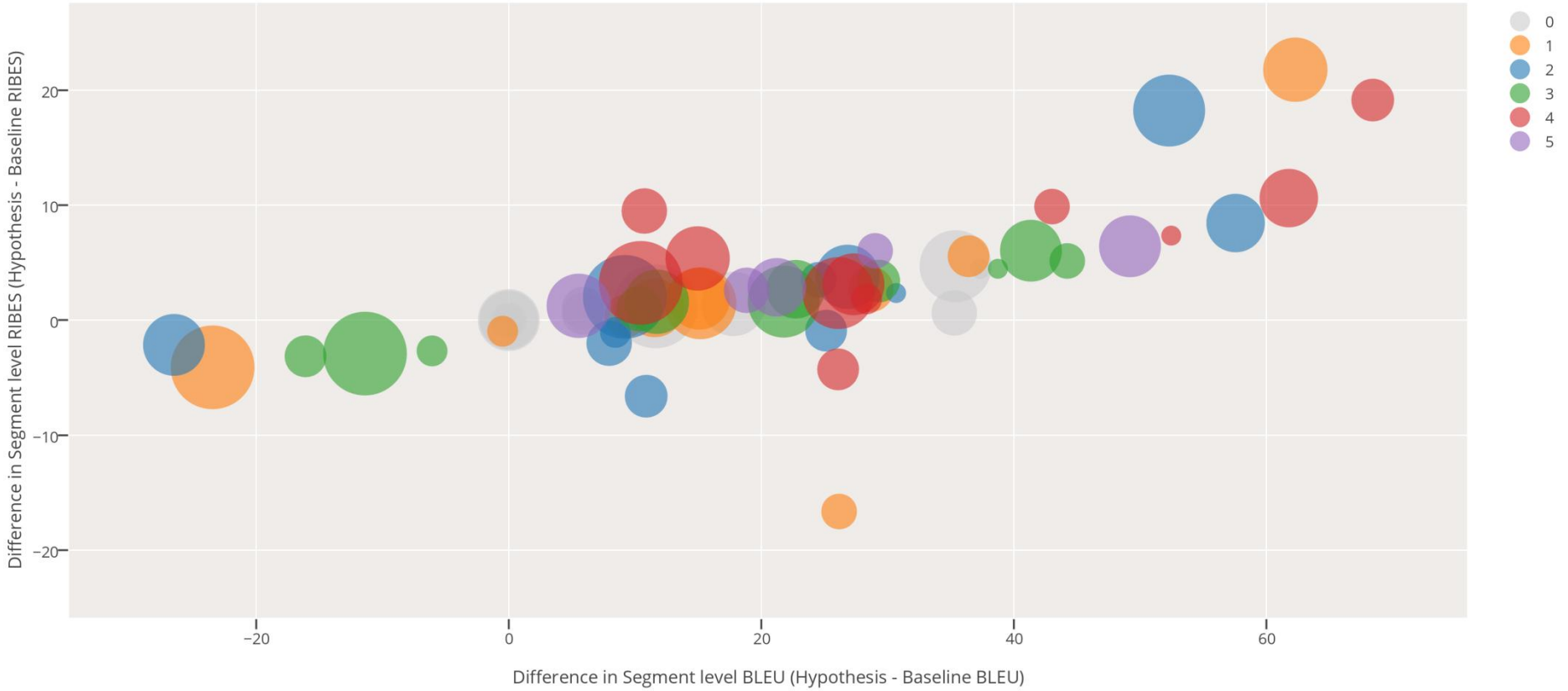
Results

(Our WAT Submission)

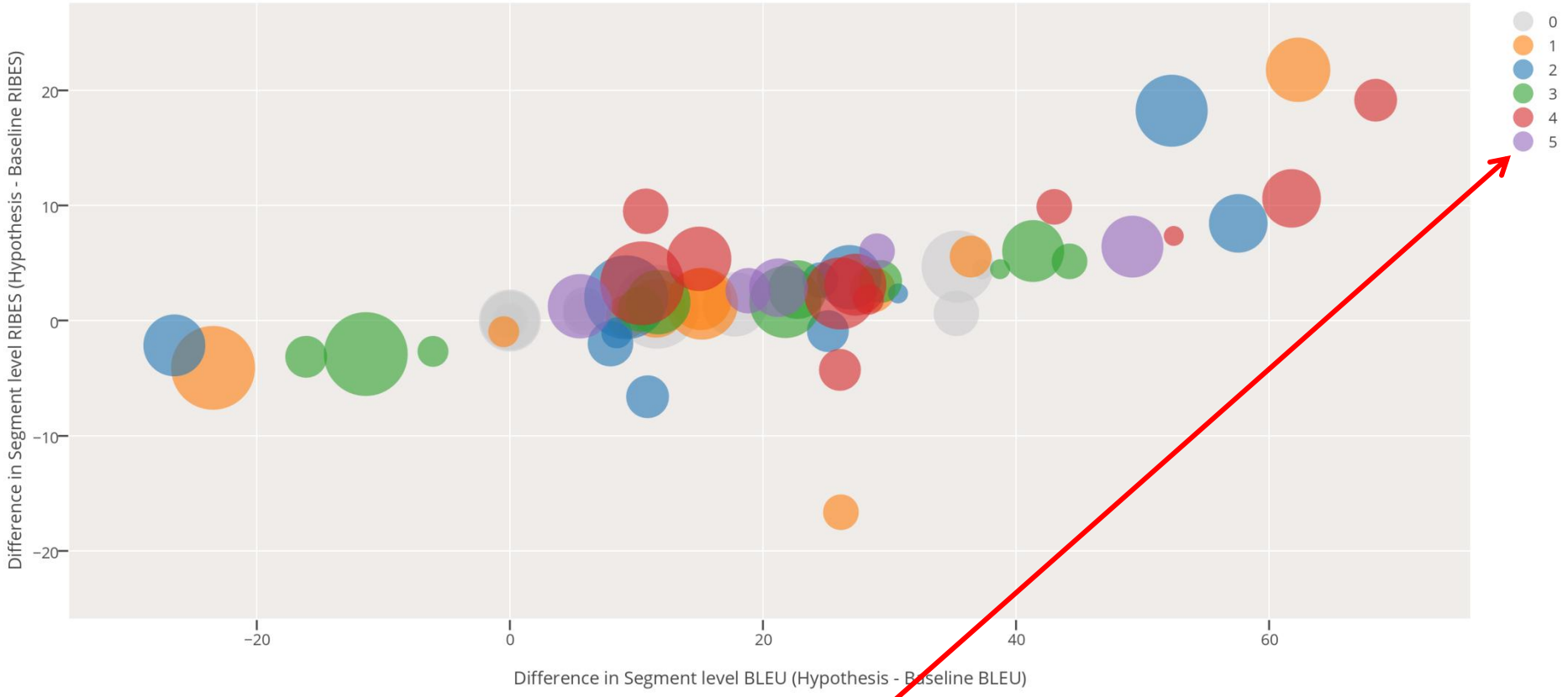
Systems	RIBES	BLEU	HUMAN
Organizers'			
PBMT baseline	94.13	69.22	0.0
Our replica			
baseline	94.29	70.23	**+3.50**
Ours (MERT 1)	95.03	84.26	-
Ours (MERT 2)	**95.15**	**85.23**	-17.75

higher BLEU = better translation is **not always true**.

Segment level Meta-Evaluation *(+ve HUMAN)*

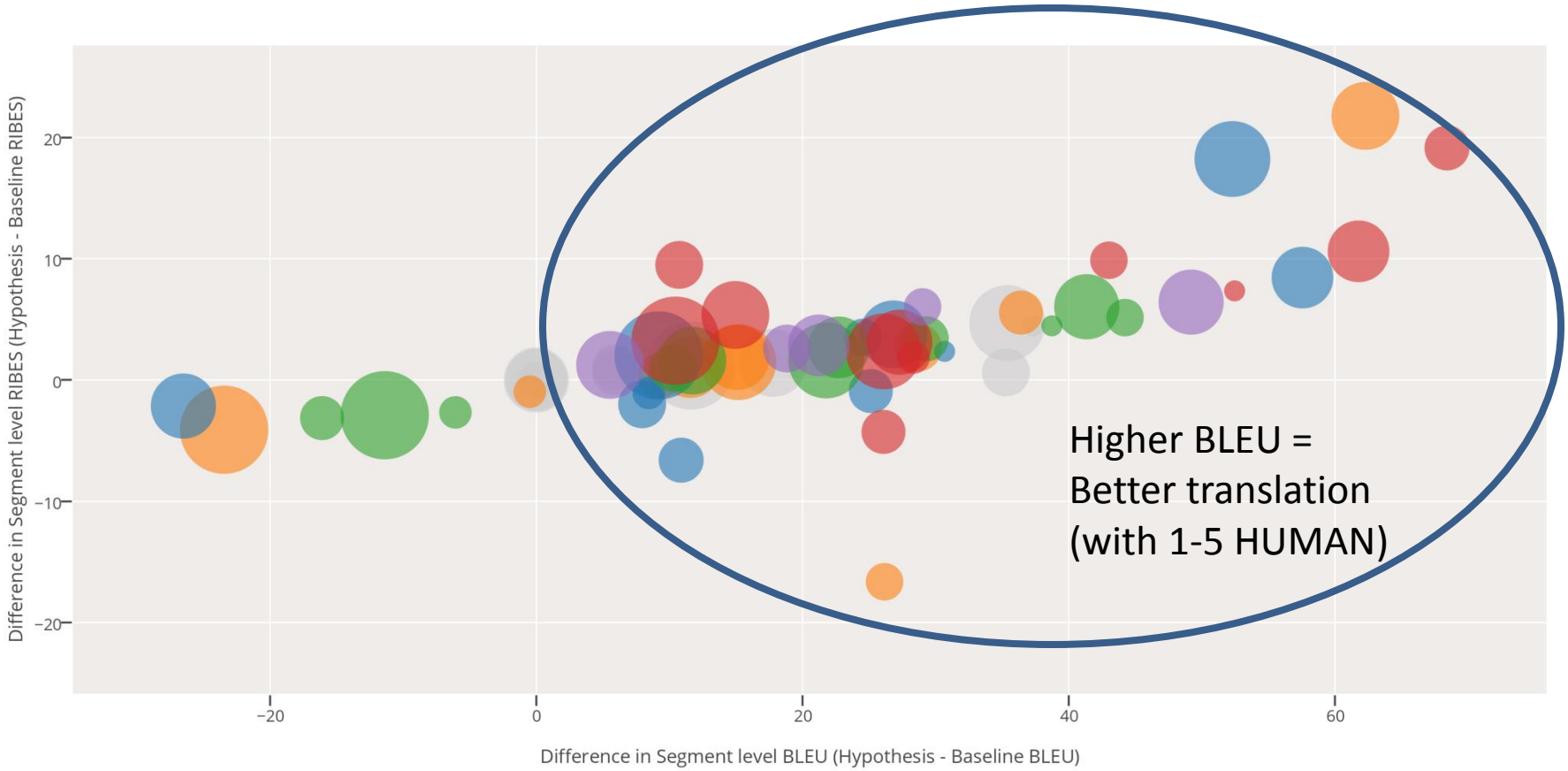


Segment level Meta-Evaluation *(+ve HUMAN)*

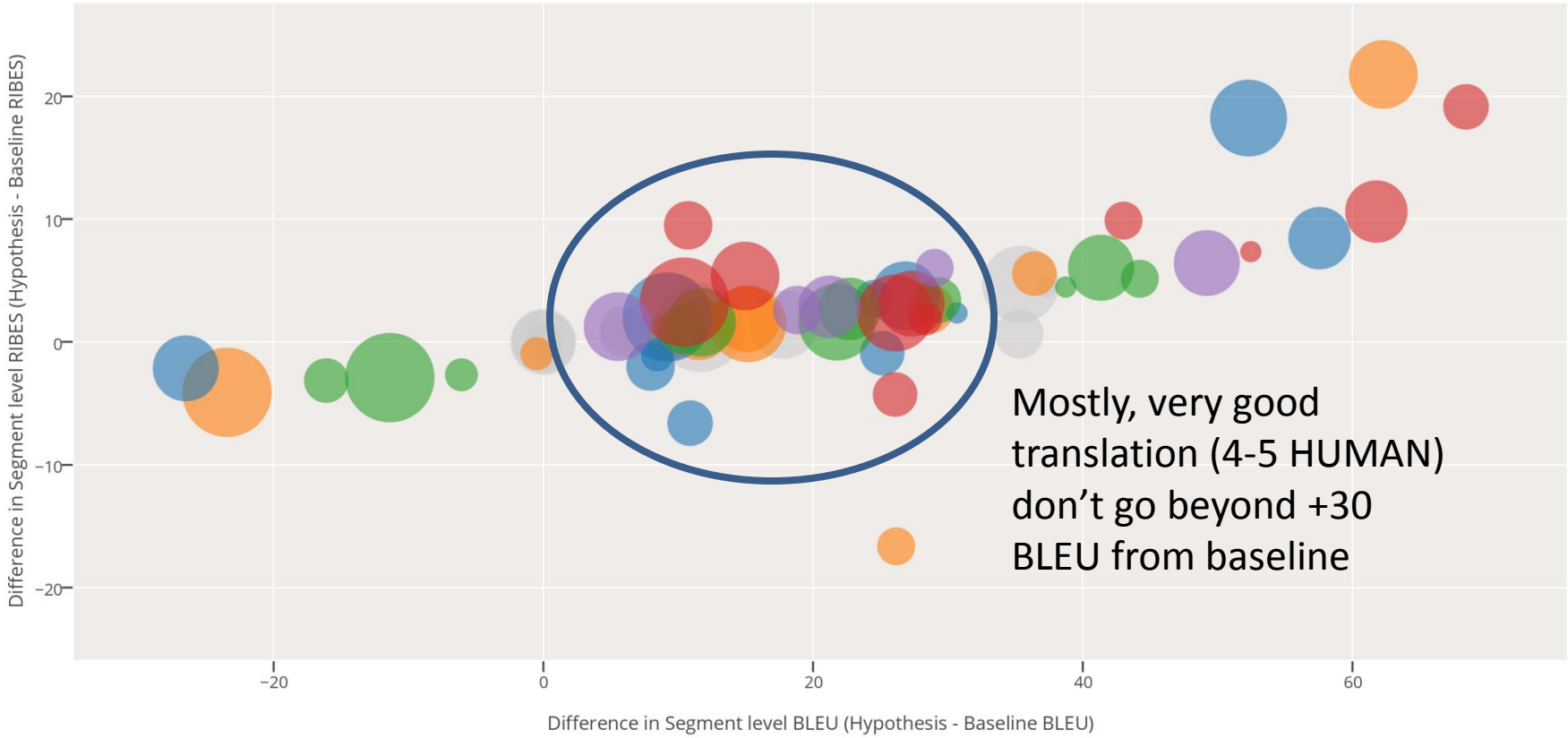


An interactive graph can be found here: https://plot.ly/171/~alvations/
(Hint: click on the bubbles here on the interactive graph

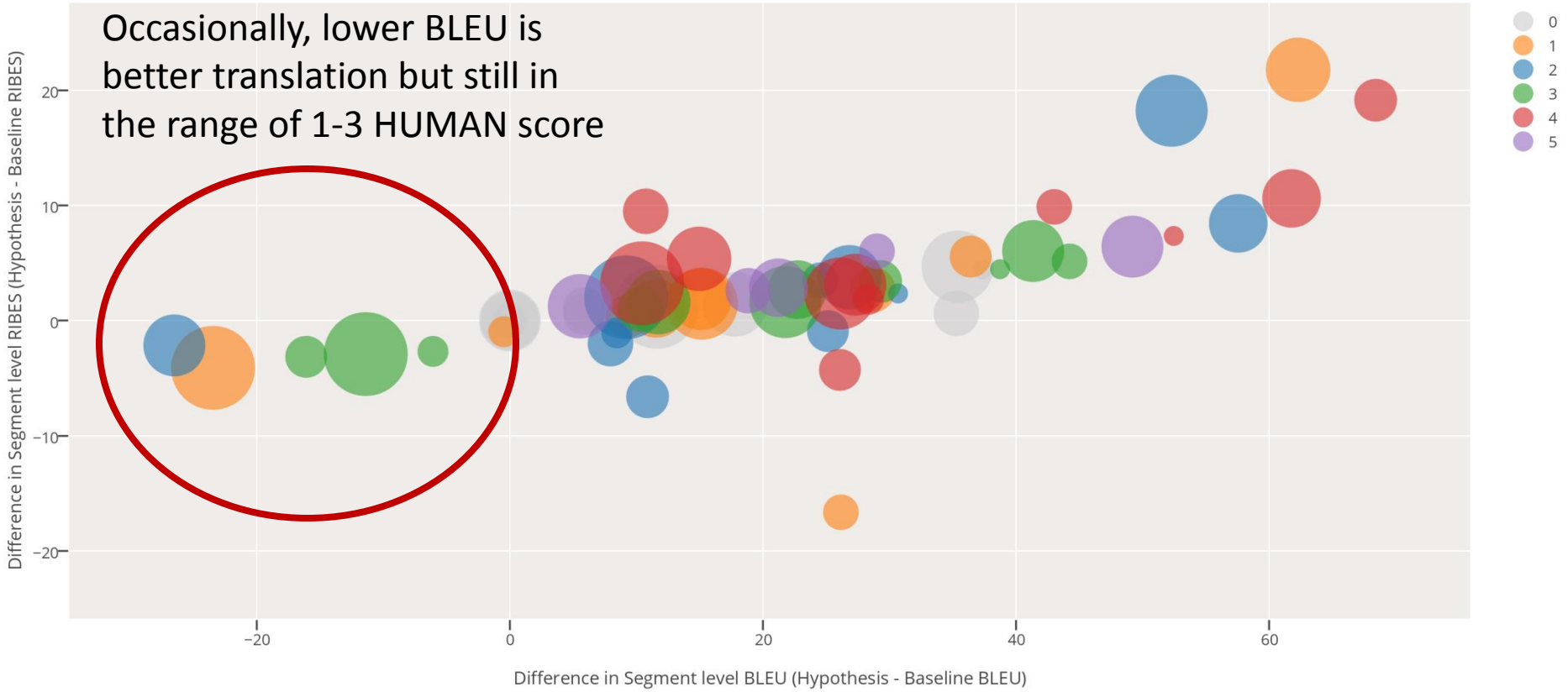
Segment level Meta-Evaluation *(+ve HUMAN)*



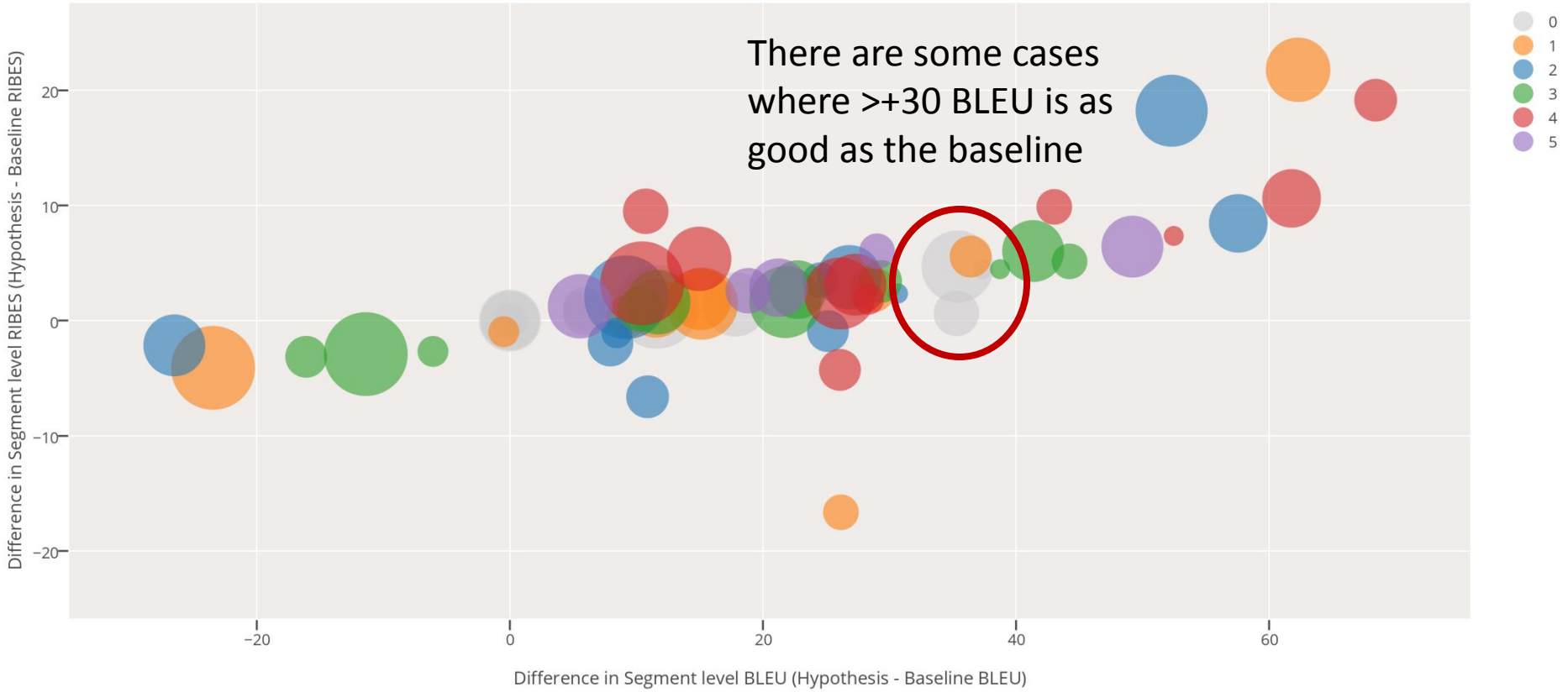
Segment level Meta-Evaluation *(+ve HUMAN)*



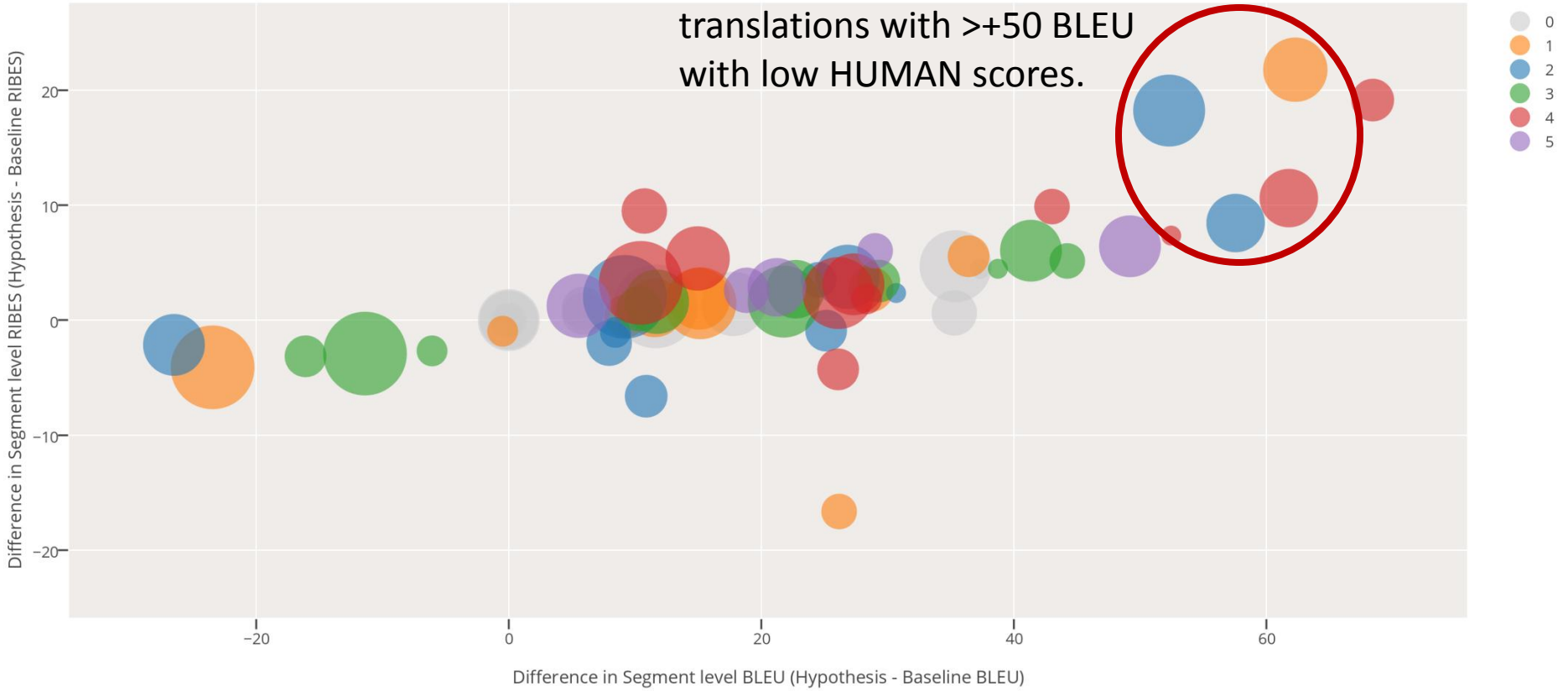
Segment level Meta-Evaluation *(+ve HUMAN)*



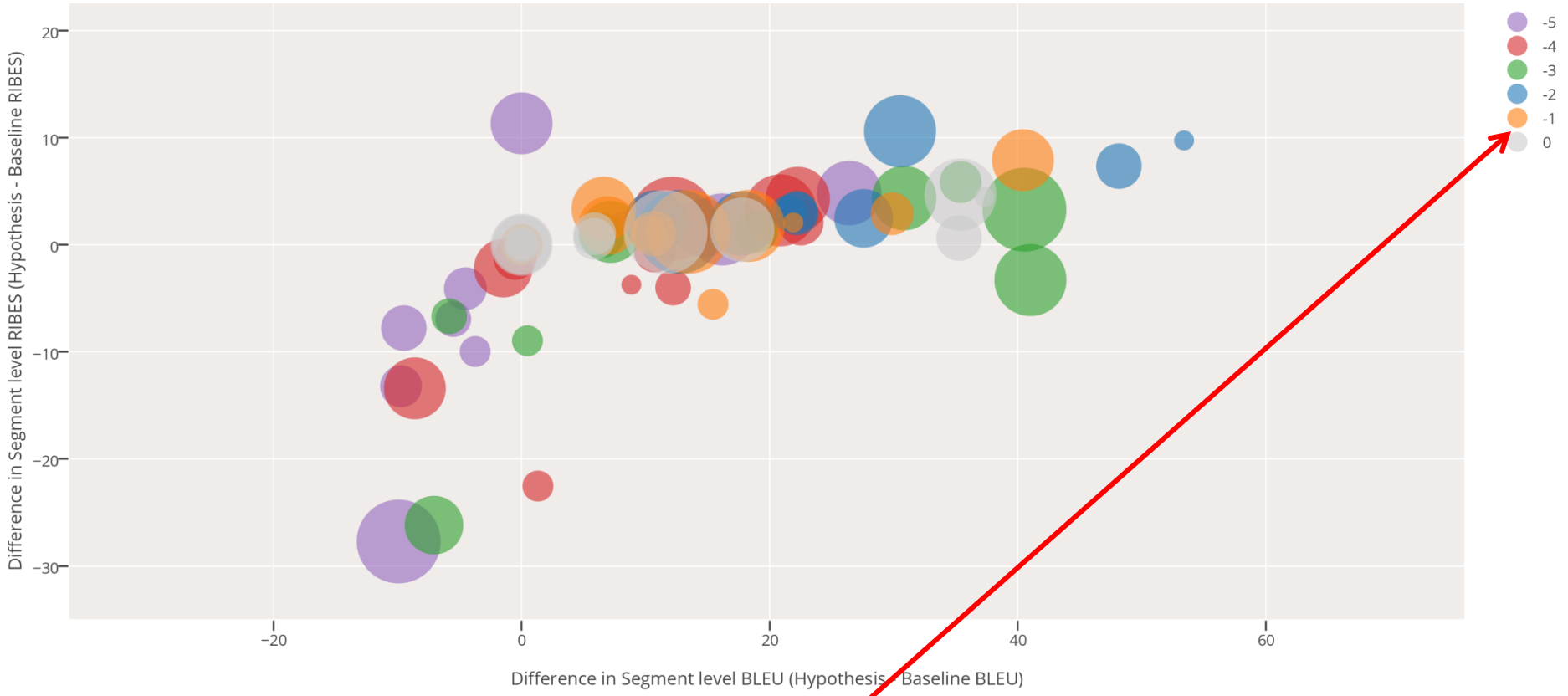
Segment level Meta-Evaluation *(+ve HUMAN)*



Segment level Meta-Evaluation *(+ve HUMAN)*

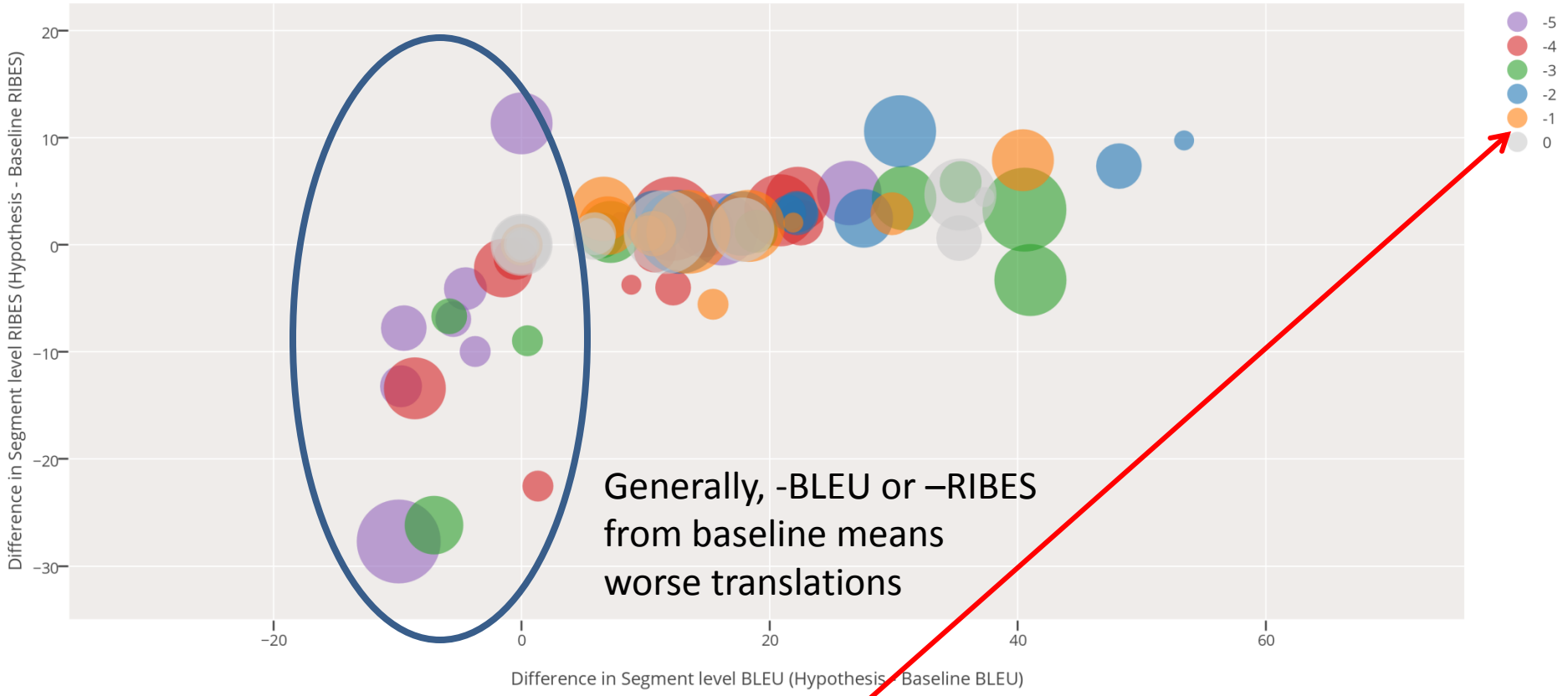


Segment level Meta-Evaluation *(-ve HUMAN)*



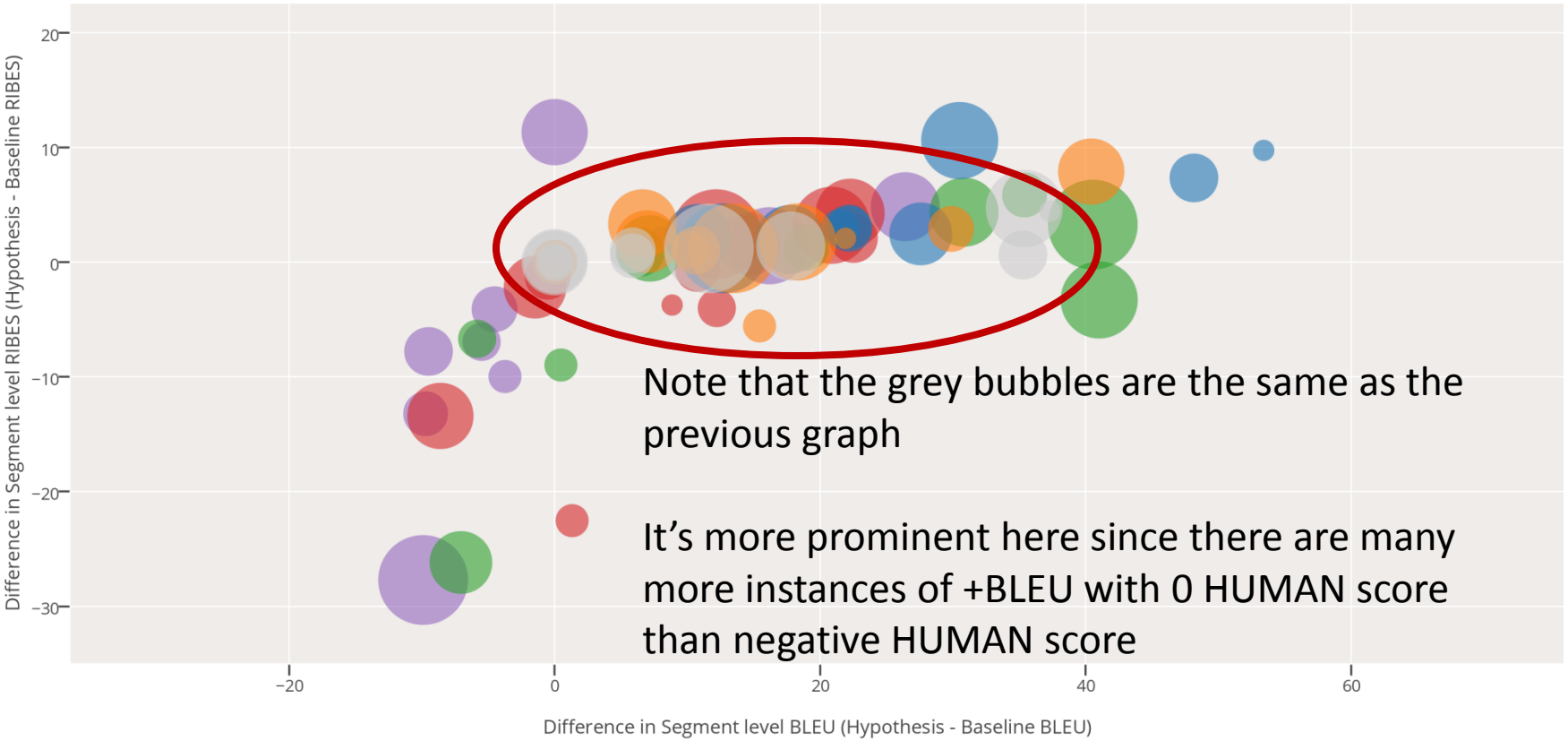
An interactive graph can be found here: https://plot.ly/173/~alvations/
(Hint: click on the bubbles here on the interactive graph

Segment level Meta-Evaluation *(-ve HUMAN)*



An interactive graph can be found here: https://plot.ly/171/~alvations/
(Hint: click on the bubbles here on the interactive graph

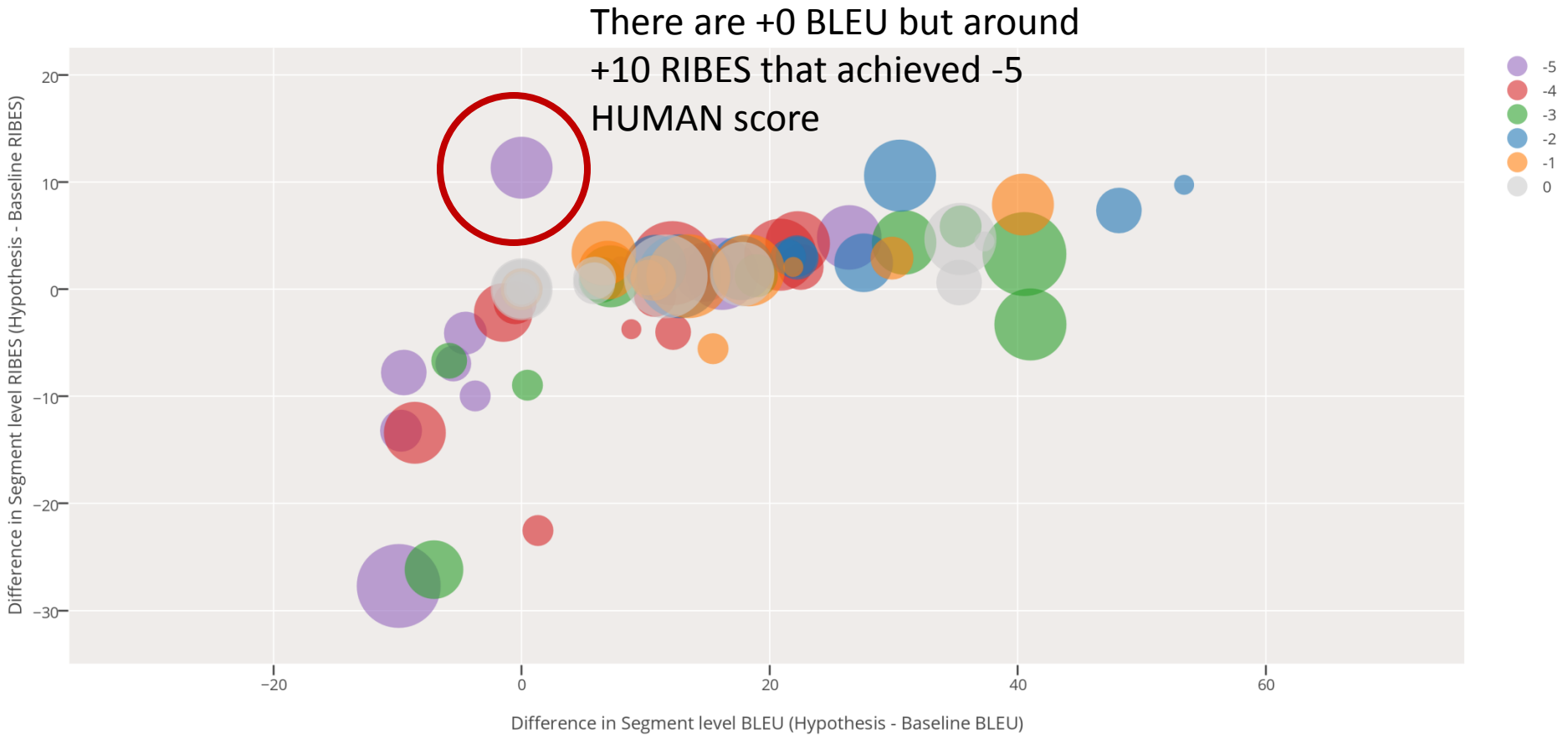
Segment level Meta-Evaluation *(-ve HUMAN)*



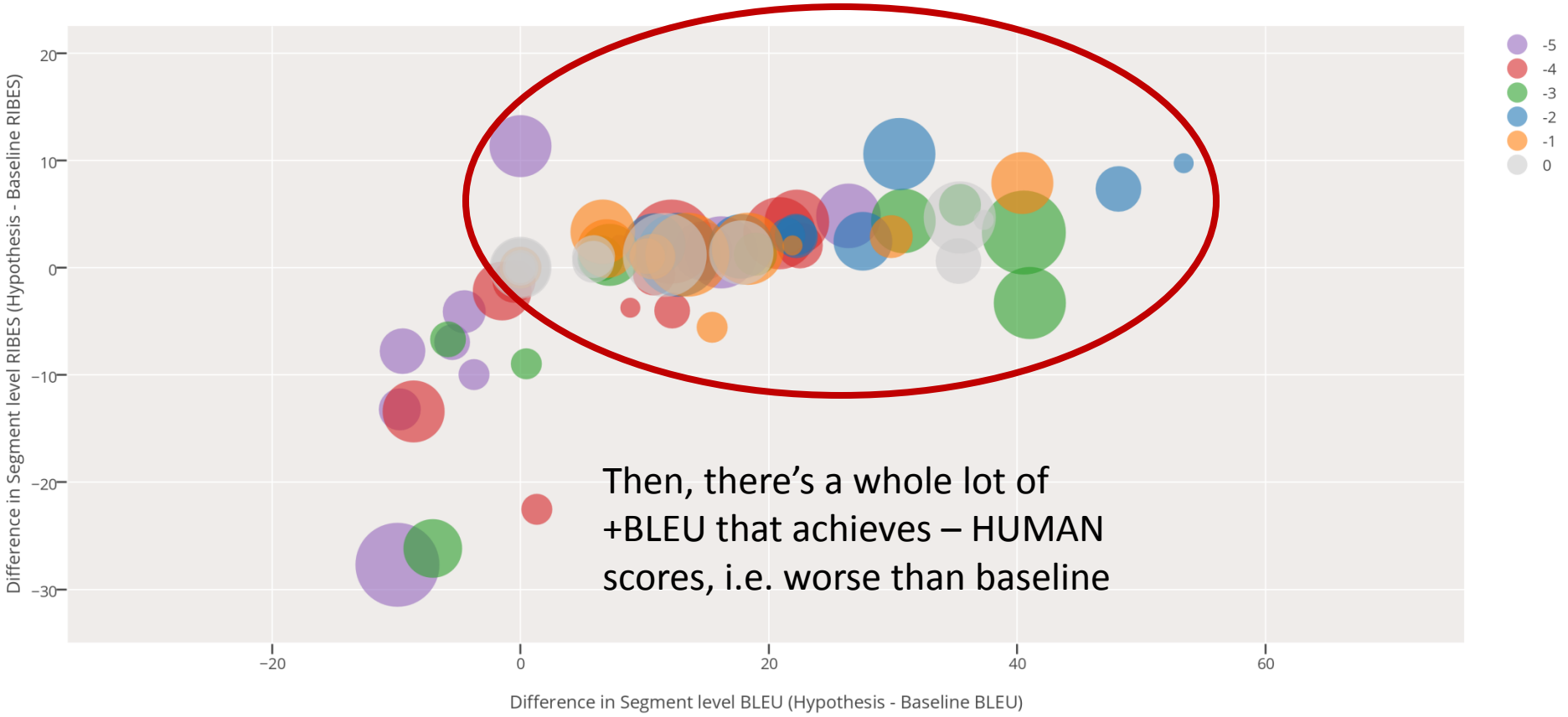
Note that the grey bubbles are the same as the previous graph

It's more prominent here since there are many more instances of +BLEU with 0 HUMAN score than negative HUMAN score

Segment level Meta-Evaluation *(-ve HUMAN)*



Segment level Meta-Evaluation *(-ve HUMAN)*



Segment level Meta-Evaluation

- With regards to positive HUMAN scores, it fits the "conventional wisdom" that
  - ***lower BLEU/RIBES = worse translation***
  - ***Higher BLEU/RIBES = better translation***

- When it comes to negative HUMAN scores, it is inconsistent with the "*conventional wisdom*"

Conclusion

- **Higher BLEU and RIBES doesn't necessary mean better translations**
  - At segment level, >+30 BLEU might not be reliable

- **Possible reasons for BLEU/RIBES to not correlate with human judgments includes:**
  - Minor lexical differences -> huge difference in n-gram precision
  - Minor MT evaluation metric differences not reflecting major translation inadequacy

References

- Bogdan Babych and Anthony Hartley. 2004. Ex- tending the BLEU MT evaluation method with frequency weightings. In ACL.
- Onderej 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 dictionary in statistical machine translation. In HyTra.

Fin.

Experiment Setup

(Our WAT Submission)

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

Results

(Our WAT Submission)

Systems	RIBES	BLEU	HUMAN
Organizers'			
PBMT baseline	94.13	69.22	0.0
Our replica			
baseline	94.29	70.23	**+3.50**
Ours (MERT 1)	95.03	84.26	-
Ours (MERT 2)	**95.15**	**85.23**	-17.75

+15 BLEU -> -17.75 HUMAN !!!

Models' Log-Linear Weights

(Our Baseline Replica)

```
# core weights
[weight]
LexicalReordering0= 0.0316949 0.0566969 0.0546839 0.0814468
0.0359473 0.0426681
Distortion0= 0.0445616
LM0= 0.274422
WordPenalty0= -0.132106
PhrasePenalty0= 0.0733761
TranslationModel0= 0.110846 0.030776 -0.013284 0.0174904
UnknownWordPenalty0= 1
```

.

Models' Log-Linear Weights

(Our MERT Run 2)

```
# core weights
[weight]
LexicalReordering0= 0.0156288 -0.0580331 0.0126421 0.0664739
0.137966 0.0303402
Distortion0= 0.048086
LM0= 0.301798
WordPenalty0= -0.029068
PhrasePenalty0= 0.0512106
TranslationModel0= 0.173756 0.0386685 -0.0237588 0.0125696
UnknownWordPenalty0= 1
```

Despite the model differences, the results shows that ***higher BLEU = better translation*** is **not always true**.