A CALL FOR CLARITY IN REPORTING BLEU SCORES

Not so long ago in a galaxy you may in fact recognize...

Everyone is excited about the new paper from some university.

I've got to add this to our notebook.

Including our hero, Stella Chercueivo!

One month later.
The new system is better, and Stella had new ideas that improved it further still!

But trouble brews. She can't match things score.

An email to the authors helps.

Meanwhile...

Stella would like to include their scores, but unfortunately, a comparison is not possible.

Curiosity moves her to investigate.

The variance in scores is larger than the gains reported in many papers.

Configuration BLEU

(2, 2) 31.2
(3, 3) 31.7
(a, 3) 31.9
none 30.0

Frustration moves her to curse!

Sacre Bleu!!

Why are you?

You can make Bleu scores comparable?

Who is the solution to all your problems?

Well, not just ours, but for future work, I suggest everyone perform different preprocessing and use the same reference.

How do we decide which one?

Typically this is accomplished with wars, but everyone is afraid of Zoetech.

Affiliation: Johns Hopkins University

Summary

\[ \text{Sacre Bleu} \]

Installation/usage are easy

\$ pip install sacrebleu

\$ cat out.detok

I can live with this.

I'll just recompute these new bests minus now. Where did I put those references?

Actually, why don't you look me up again.

Okay... thanks!

Even better my real goodness.

Even better my real goodness.

The grief you may save is your own.

Work done while at: Amazon

Code available at: github.com/mjpost/sacrebleu