A CALL* FOR CLARITY IN REPORTING BLEU SCORES

*AND PROPOSAL

MATT POST

NOT SO LONG AGO IN A GALAXY YOU MAY IN FACT RECOGNIZE....

EVERYONE IS EXCITED ABOUT THE NEW PAPER FROM ACME UNIVERSITY.

FRITTER

ACME: "Introducing the new Transmogrifier model...."

"whoa so much bleu score"

"This reminds me of when I had this idea in 1997 http://arxiv.org/..."

I've got to add this to our toolkit.

INCLUDING OUR HERO, STELLA CHERCHEURE!

SHE GETS TO WORK.

$P(f|e) = P(e|f)P(f)$

$y = \text{softmax}(Wx+$

~ONE MONTH LATER~

THE NEW SYSTEM IS BETTER, AND STELLA HAD NEW IDEAS THAT IMPROVED IT FURTHER STILL!

I can't believe no one thought to use Lagrange multipliers and Kronecker delta functions before.

BUT TROUBLE BREWS. SHE CAN'T MATCH ACME'S SCORES.

SYSTEM	THEM	ME
baseline	x	y
Transmogrifier	x+3	y+3
+ new ideas	—	y+4

x ? y

PROBLEM: BLEU IS UNDERSPECIFIED—AND DETAILS CAN BE HARD TO FIND IN PAPERS.

AN EMAIL TO THE AUTHORS HELPS

FROM: AUTHOR 2
DATE: A WEEK LATER
SUBJECT: Re: (not) feelin' blue

Hi Stella,

In the rush to meet the ArXiv deadline, we forgot to mention we applied $(\lambda(\lambda))$-pre-processing...

WITH THIS INFO, SHE ESTABLISHES PARITY. TIME TO WRITE THINGS UP FOR *ACL.

MEANWHILE....

DIVERGE

TECH — SCIENCE — SENSATIONALISM

INITECH'S NEW TRANSLATOR INVENTS ITS OWN AI!! - ARE NUCLEAR WEAPONS NEXT?

HOT STARTUP INITECH HAS SOMETHING EVEN NEWER!

STELLA WOULD LIKE TO INCLUDE THEIR SCORES, BUT UNFORTUNATELY, A COMPARISON IS NOT POSSIBLE.

$(\beta(\lambda(\xi))$-pre-processing?! You have got to be kidding me.

PROBLEM: DIFFERENT REFERENCE PROCESSINGS CANNOT BE COMPARED.

CURIOSITY MOVES HER TO INVESTIGATE

UNKS

casing?

compound splitting?

reference count

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

tokenization

test set

Configuration	BLEU
$(\lambda(\lambda))$	31.2
$(\beta(\lambda(\xi))$	31.7
$(\alpha(\omega))$	31.4
none	33.0

FRUSTRATION MOVES HER TO CURSE!

SACRE BLEU!! ಠ_ಠ

Did someone call my name?

Who are you?

I'm the solution to all* your problems!

You can make BLEU scores comparable?

Well, not past ones, but for future work, I suggest everyone scores detokenized outputs with the same reference preprocessing

How do we decide which one?

Typically this is accomplished with wars, but everyone's afraid of Initech...

*CLOSED WORLD ASSUMPTION

A neutral choice would be to use WMT's. They provide many of our test sets, after all.

I can live with that.

I'll just recompute those numbers a minute now. Where did I put those references?

Actually, why don't you let me handle that?

Okay... thanks!

SOLUTION: WHEN REPORTING SCORES, USE WMT SCORING ON DETOKENIZED SYSTEM OUTPUTS.

SUMMARY

- BLEU SCORES CAN VARY WILDLY WITH DIFFERENT PARAMETERIZATIONS
- PAPERS OFTEN DO NOT REPORT ALL THE DETAILS
- THIS IS A NEEDLESS IMPEDIMENT TO SCIENCE!

TWO PROPOSED REMEDIES:

1. USE WMT SCORING (ON DETOKENIZED REFERENCES) — SO SCORES CAN BE DIRECTLY CMP'D ACROSS PAPERS
2. INCLUDE ALL THE DETAILS IN THE WRITEUP

SACREBLEU CAN HELP!

SACREBLEU

INSTALLATION + USAGE ARE EASY

```
$ pip install sacrebleu
$ cat out.detok \
    | sacrebleu -t wmt18 -l de-en
```

- SCORES COMPARABLE W/ MATRIX.STATMT.ORG
- DOWNLOADS COMMON REFERENCES FOR YOU
- INFO STRING TO SAVE YOUR PEERS SOME GRIEF!

BLEU+c.mixed+l.de-en+tok.13a+#.1+v.1.2.12

EVEN BETTER MT

Stella Cherchcheure

WITH COMPARABLE SCORES!

THE GRIEF YOU SAVE MAY BE YOUR OWN.

AFFILIATION: JOHNS HOPKINS UNIVERSITY

WORK DONE WHILE AT: amazon

CODE AVAILABLE AT:

github.com/mjpost/sacrebleu