Holy and Unholy Grails

Eduard Hovy
and Deepak Ravichandran
USC Information Sciences Institute
Seeker for the Holy Grail, beware…

when you find it, it might turn out to be less holy than you wanted…

and then you’re stuck with a problem.

Third Book of Yorickus, verse 4
Automated MTE: BLEU

• Proposed by Papineni et al. 2001
  – Create $N$ human translations (ideals)
  – Get system translations
  – Count number of ngram overlaps (1gram, 2, 3, 4)
  – Combine them, with length penalty:
    \[ \text{BLEU} = LP(a_{1.1gr} + a_{2.2gr} + a_{3.3gr} + a_{4.4gr}) \]
  – Observed very good correspondence with people’s relative rankings (to 3rd dec place!)

• Compare to human score: human = 100%, system = ?
It’s too easy to look good

Hindi-English system, June 2003

Explosion for kept

You view will be 23 June 1985 to Air India a number 182 Canada a New Delhi for which lasted was but did over bomb explosion was which all 329 passengers were killed. Now these that the plane bomb explosion in a major accused Singh were their own crime accepted that he only that in the plane bomb were kept. Singh were of statement after they Canada a court the five years of of were was heard.

BLEU score: 73% of human (avg)
Some examples

**Ideal**: George Bush will often take a holiday in Crawford Texas

|                           | BLEU  
|---------------------------|-------
| George Bush will often take a holiday in Crawford Texas | 1.0000 |
| Bush will often holiday in Texas                          | 0.4611 |
| Bush will often holiday in Crawford Texas                  | 0.6363 |
| George Bush will often holiday in Crawford Texas           | 0.7490 |
| George Bush will not often vacation in Texas               | 0.4491 |
| George Bush will not often take a holiday in Crawford Texas| 0.9129 |
BLEU, the Holy Grail of MTE

Which is better?

George Bush often takes a holiday in Crawford Texas
holiday often Bush a takes George in Crawford Texas

What would BLEU say?

If the human’s ‘gold standard’ comparison is:

The President frequently makes his vacation in Crawford Texas

Then the scores are…

0.2627 0.2627
Why? The invisible region

Ideal: The President frequently makes his vacation in Crawford Texas

S1.1: George Bush often takes a holiday in Crawford Texas

S1.2: holiday often Bush a takes George in Crawford Texas

Ideal: In Paris, she and Nicholas said goodbye before she got on the train and went to Cologne

S2.1: In Paris, she and Nicholas said goodbye before she [got on the train] and went to Cologne

S2.2: In Paris, she [got on the train] and Nicholas said goodbye before she [] and went to Cologne

BLEU 4g

0.2627

0.2627

BLEU 2g

1.0

1.0
Fixin’ da blues (here in New Orleans…)

H1.0: The President frequently makes his vacation in Crawford Texas
S1.1: George Bush often takes a holiday in Crawford Texas
S1.2: holiday often Bush a takes George in Crawford Texas

POS only:
H1.0: DT NNP RB VBZ PRP$ NN IN NNP NNP
S1.1: NNP NNP RB VBZ DT NN IN NNP NNP
S1.2: NN RB NNP DT VBZ NNP IN NNP NNP

(Words + POS) / 2:
S1.1:
S1.2:

BLEU 4gram

<p>| | | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2627</td>
<td>0.2627</td>
<td>0.5411</td>
<td>0.3217</td>
<td>0.4020</td>
<td>0.2966</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Compare BLEU & BLEU+ scores for pairs (human-system)
Some quotes

• “Yes, SMT is the Holy Grail—but it will take 20 or 200 years to reach” — Hermann Ney
• “At sufficient distance, anything looks like a Grail” — Epistle of Sergeius 14:3

• “Have we reached the Holy Grail?”
• SMT has been around for about 15 years
• “Ye shall know its worth when the customers pay for it” — Analect in Book of Denis
More quotes

• Observation: SMT engines are increasingly duplicating the traditional architectures:
  – Word replacement: CANDIDE 1992
  – + some ordering, morphology: CANDIDE ~1995
  – Pre-syntactic local trees: Wu ~1995
  – + some syntax: Yamada and Knight 2001
  – + EBMT phrase patterns: Och 2003

• “When you execute the same steps using a different tool, you have invented a new tool, not a new world” — from St. Jamius’s stele
So don’t get happy too soon...

...what looks like gold might be tin

(Epistle of Ed)