MT
Evaluating Systems

Miles Osborne

School of Informatics
University of Edinburgh
miles@inf.ed.ac.uk

January 11, 2010

Need for evaluation

Central to progress in any field is good evaluation

- How do we know if we are doing a good job?
- What is wrong with the current system?
- Are we the best in the world?

Evaluating MT is hard . . .

MT Evaluation

What counts as a good translation?

The cat sat on the mat	(reference)
The cat sat on mat the	1
On the mat sat the cat	2
The cat on the floor	3
A cat sat on the mat	4
the cat sat on the mat	5
The cat sat on the straw mat	6

Subjective Evaluation

One could argue that a good translation is one that people like:

The cat sat on the mat	(reference)
The cat sat on the floor	1
On the mat sat the cat	2
The cat on the floor	3
A cat sat on the floor	4
the cat sat on the mat	5
The cat sat on the straw mat	6

Subjective Evaluation

One approach is to have people rate candidates

- Which candidate sentence captures most of the meaning?
  - *Adequacy*
- Which candidate sentence is most readable?
  - *Fluency*

Subjective Evaluation

- People have difficulties carrying out this task:
  - There is a preference for fluency.
  - People tend not to agree with each other.
- It is time consuming.
- Subjective evaluation is hard to reuse.

Subjective Evaluation

An alternative is to rank sentences:

- Do you prefer sentence A to sentence B?

Sentence ranking is faster and more reliable than sentence rating

Automatic Evaluation

Since people are expensive, *automatic* methods have become popular

- Create a set of *reference* translations.
- Design a *similarity metric* to measure sentence closessness.
- Apply that metric between the candidate and the set of references.

There are many automatic methods

First attempt

- We want to be rewarded if we produce output that is present within a reference.
  - This aspect deals with word choice.
- We should be rewarded if that output is in the correct order.
  - This aspect deals with word order.

First attempt

Count ngrams appearing in the candidate and in a reference

- An ngram of order one (unigram) rewards word choice.
- Higher order ngrams reward word order.
- (Four is a useful maximum order)

The cat sat	(reference)
The cat	(The cat) (The) (cat)
sat the cat	(sat) (the) (cat)

Counting these ngrams tells us what we have found

First attempt

Count ngrams appearing in the reference but not in the candidate

The cat sat	(reference)
The cat	(The cat sat) (cat sat) (sat)
sat the cat	(The cat) (The) (The cat sat)

Counting these ngrams tells us what we have missed

First attempt

Putting it together:

- Sum-up the ngram hits (order n) in candidate translation c_i
  - Call this h_{c_i}
- Sum-up the ngram hits for all candidate translations.
- Sum-up the total number of possible order n ngrams in reference r
  - Call this t_{r_i}
- Sum-up the total possible hits over all reference translations.

$$p_n = \frac{\sum_i h_{c_i}}{\sum_i t_{r_i}}$$

Second attempt

We can *game* the metric:

- Produce lots of function words.
- Produce all possible ngrams!
  - The The The The cat cat cat cat ...
- Counts are *clipped* to prevent double counting.

Second attempt

How well does this do?

- IBM took output from a human translator.
- IBM also took output from a poor machine translator.

Ngram order	Human	Machine
1	.8	.5
2	.5	.1
3	.3	.05
4	.2	.01

$\longrightarrow$ Easy to tell people from machines

Third attempt

We do not deal with candidate translations that are too short

- The *Brevity Penalty* (BP) punishes translations that are too short.
- The BP is document-based.
- Ngram hits punish translations if they are too long (produce spurious ngrams)

Third attempt

The BP:

$$BP = \begin{cases} 1 & \text{if } C' \geq R' \\ \exp(1 - \frac{R'}{C'}) & \text{if } C' < R' \end{cases}$$

Here R' is the total length (in words) of all selected references (ie the document) and C' is the sum of all candidate translations (ie the translated document)

Third attempt

$$BP = \begin{cases} 1 & \text{if } C' \geq R' \\ \exp(1 - \frac{R'}{C'}) & \text{if } C' < R' \end{cases}$$

- By construction the BP is a probability.
- As translations get shorter the BP decreases

C'	R'	BP	
5	10	0.37	Too short
8	10	0.78	Too short
10	10	1	Same length
12	10	1	Too long

Putting it together

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

- The first part is our brevity penalty
- The second part is a geometric mean of ngram matches
- (Each match is weighted if for example we prefer four-gram hits)

This is the *BLEU* metric and it is widely used in MT research

Comments on BLEU

- BLEU generally correlates with subjective evaluation.
- It is hard to interpret.
- BLEU cannot be used across language pairs.
- It should be used with caution since good translations not in the reference set are punished

Summary

- Subjective evaluation is the best way to measure progress
- Automatic metrics are useful for day-to-day evaluation
- Care must be taken interpreting BLEU scores