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 . . .
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

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.

One approach is to have people rate candidates:
- Which candidate sentence captures most of the meaning?
  - Adequacy
- Which candidate sentence is most readable?
  - Fluency

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 closeness.
- Apply that metric between the candidate and the set of references.

There are many automatic methods

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)

\[
\text{The cat sat} \quad \text{(reference)} \\
\text{The cat} \quad \text{(The cat) (The) (cat)} \\
\text{sat the cat} \quad \text{(sat) (the) (cat)}
\]

Counting these ngrams tells us what we have found

First attempt

Count ngrams appearing in the reference but not in the candidate

\[
\text{The cat sat} \quad \text{(reference)} \\
\text{The cat} \quad \text{(The cat) (cat) (sat)} \\
\text{sat the cat} \quad \text{(The) (The cat sat) (The cat sat)}
\]

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 $t_r$
- Sum-up the total possible hits over all reference translations.

$$p_n = \frac{\sum_{i} h_{c_i}}{\sum_{i} t_r}$$

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.

How well does this do?

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

<table>
<thead>
<tr>
<th>Ngram order</th>
<th>Human</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.8</td>
<td>.5</td>
</tr>
<tr>
<td>2</td>
<td>.5</td>
<td>.1</td>
</tr>
<tr>
<td>3</td>
<td>.3</td>
<td>.05</td>
</tr>
<tr>
<td>4</td>
<td>.2</td>
<td>.01</td>
</tr>
</tbody>
</table>

→ 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 (i.e., the document) and \( C' \) is the sum of all candidate translations (i.e., the translated document).

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.

The BP is a probability. As translations get shorter the BP decreases.

\[
\begin{array}{ccc}
C' & R' & BP \\
5 & 10 & 0.37 \text{ Too short} \\
8 & 10 & 0.78 \text{ Too short} \\
10 & 10 & 1 \text{ Same length} \\
12 & 10 & 1 \text{ Too long}
\end{array}
\]

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

- The first part is our brevity penalty
- The second part is a geometric mean of n-gram 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.
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