1 Evaluation of SMT systems: BLEU

Idea: We want to define a repeatable evaluation method that uses:

- a gold standard of human generated **reference translations**
- a numerical **translation closeness** metric in order to compare the system output against human references

Advantages: Less costy than repeated manual evaluation.
Central idea of BLEU: We define "translation closeness" by counting matches of **n-grams** in candidate and reference translation.

1.1 Modified n-gram precision

Example: Candidate: the the the the the the the
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

Unigram precision: $\frac{7}{7} = 1$
Modified unigram precision: $\frac{2}{7}$.

1.1.1 Modification: Clipping + corpus-based calculation

1. For each n-gram, for each candidate sentence, count the maximal number of n-gram matches in a single reference translation.
2. For each n-gram, for each candidate sentence, clip the total number of matches of a candidate n-gram by the maximal reference match.
3. For each n-gram, add up clipped matches over all candidate sentences in corpus.
4. For each n-gram, divide by the total number of unclipped candidate n-gram counts in corpus.

$$p_n = \frac{\sum_{c \in \{\text{candidates}\}} \sum_{\text{n-gram} \in c} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{c' \in \{\text{candidates}\}} \sum_{\text{n-gram}' \in c'} \text{count}_(\text{n-gram}')} \tag{1}$$

Example
Candidate: of the
Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed the directions of the Party.

$$\text{Modified unigram } p = \frac{1 \cdot \text{of} + 1 \cdot \text{the}}{1 \cdot \text{of} + 1 \cdot \text{the}} = \frac{2}{2} = 1 \tag{2}$$

$$\text{Modified bigram } p = \frac{1 \cdot \text{of the}}{1 \cdot \text{of the}} = \frac{1}{1} = 1 \tag{3}$$

1.1.2 Combining modified n-gram precisions

Unigram precision is exponentially larger than bigram precision, etc.
Exponential decay is formalized as the **weighted average** of log n-gram precisions. This is equivalent to the **log of geometric mean**.

$$\sum_{n=1}^{N} \frac{1}{N} \log p_n = \log \underbrace{(\prod_{n=1}^{N} p_n)^{\frac{1}{N}}}_{\text{geometric mean}} \tag{4}$$

1.2 The trouble with recall

1.2.1 Problem: Recalling more words is worse

Example:
Candidate 1: I always invariably perpetually do.
Candidate 2: I always do.
Reference 1: I always do.
Reference 2: I invariably do.
Reference 3: I perpetually do.

Recall(candidate 1) > Recall(candidate 2); does not make sense! Recalling all choices from multiple references leads to bad translations.

1.2.2 Alternative: Brevity penalty

Too long candidates are penalized by n-gram precision, while too short candidates are penalized by brevity penalty.
Important: Brevity penalty is computed over corpus in order to avoid harsh penalties on short sentences.

The **corpus reference length** r is the sum over best match lengths for each candidate (or shortest if equally close match).
The **corpus candidate length** c is the total length of candidates in corpus.
Brevity penalty decays exponential in $\frac{r}{c}$.

1.3 BLEU metric

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

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ \exp(1 - \frac{r}{c}) & \text{if } c \leq r. \end{cases}$$

$$\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} \frac{1}{N} \log p_n$$

BLEU Example:

Reference 1:
Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.

Reference 2:
Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.

Reference 3:
Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.

Reference 4:
Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.

Candidate:
Appeared calm when he was taken to the American plane, which will to Miami, Florida.

1-grams: **American**, **Florida**, **Miami**, Orejuela, **appeared**, as, being, **calm**, carry, escorted, **he**, him, in, led, **plane**, quite, seemed, take, that, **the**, **to**, **to**, to, **was** , was, **which**, while, **wil**l, would, **,**, **.**

2-grams: **American plane**, **Florida .**, **Miami ,**, Miami in, Orejuela appeared, Orejuela seemed, **appeared calm**, as he, being escorted, being led, calm as, calm while, carry him, escorted to, **he was**, him to, in Florida, led to, plane that, plane which, quite calm, seemed quite, take him, that was, that would, **the American**, the plane, **to Miami**, to carry, **to the**, was being, was led, was to, **which will**, while being, will take, would take, **, Florida**

3-grams: American plane that, American plane which, **Miami , Florida**, Miami in Florida, Orejuela appeared calm, Orejuela seemed quite, appeared calm as, appeared calm while, as he was, being escorted to, being led to, calm as he, calm while being, carry him to, escorted to the, he was being, he was led, him to Miami, in Florida ., led to the, plane that was, plane that would, plane which will, quite calm as, seemed quite calm, take him to, that was to, that would take, **the American plane**, the plane that, **to Miami ,**, to Miami in, to carry him, **to the American**, to the plane, was being led, was led to, was to carry, which will take, while being escorted, will take him, would take him, **, Florida .**

unigram precision $p_1 = \frac{15}{18}$, bigram precision $p_2 = \frac{10}{17}$, trigram precision $p_3 = \frac{5}{16}$

1.4 Problems with BLEU

1. BLEU is **not sufficient** to reflect genuine translation quality.
2. BLEU improvement is **not necessary** for improved translation quality.

Example for 1.:
Permutations on unigram or bigram level do not reduce BLEU. For b bigram matches in a candidate of length K there are $(K - b)!$ permutations.

Example for 2.:
Translations from radically different systems are not well distinguished by BLEU.

Fluency
How do you judge the fluency of this translation?
5 = Flawless English
4 = Good English
3 = Non-native English
2 = Disfluent English
1 = Incomprehensible

Adequacy
How much of the meaning expressed in the reference translation is also expressed in the hypothesis translation?
5 = All
4 = Most
3 = Much
2 = Little
1 = None.

Iran has already stated that Kharazis statements to the conference because of the Jordanian King Abdullah II in which he stood accused Iran of interfering in Iraqi affairs.
n-gram matches: 27 unigrams, 20 bigrams, 15 trigrams, and ten 4-grams **human scores:** Adequacy:3,2 Fluency:3,2
Iran already announced that Kharrazi will not attend the conference because of the statements made by the Jordanian Monarch Abdullah II who has accused Iran of interfering in Iraqi affairs.
n-gram matches: 24 unigrams, 19 bigrams, 15 trigrams, and 12 4-grams **human scores:** Adequacy:5,4 Fluency:5,4
Reference: Iran had already announced Kharazi would boycott the conference after Jordans King Abdullah II accused Iran of meddling in Iraqs affairs.

1.5 Translation Error Rate (TER)

An error metric for MT that measures the minimum number of edits required to change a system output into one of the references.

An edit is a elementary operation from the set:

1. word insertion
2. word deletetion
3. word substitution
4. block of words move (phrasal shift)

$$\text{TER} = \frac{\#\text{ of edits}}{\text{average }\#\text{ of reference words}}$$

Note: Unlike BLEU, lower TER scores are better.

A shift (operation 4) moves a contiguous sequence of words within the hypothesis to another location within the hypothesis. All edits, including shifts of any number of words, any distance, have *equal unit cost*.

Example:

```
REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times
HYP: THIS WEEK THE SAUDIS denied information published in the new york times
```

(capitalization added for emphasis)

- "this week" in the HYP is in a "shifted" position (at the beginning rather than after "denied") with respect to the REF.
- "Saudi Arabia" in the REF appears as "the Saudis" in the HYP (counts as 2 separate substitutions).
- "American" appears only in the REF.

Hence, the number of edits is 4 (1 shift, 2 substitutions and 1 insertion), giving a TER score of $\frac{4}{13} = \mathbf{0.13}$ (not bad). Compare to BLEU which yields a pretty low score of 0.32% (breakdown over n-grams: 0.833/0.545/0.300/0.111, brevity 0.920). To have the error-rate analog to the TER, we should compare $1-\text{BLEU}= 1 - 0.32 = \mathbf{0.68} \gg \mathbf{0.13}$.

As BLEU fails to account for phrasal shifts adequately it overly penalises the hypothesis, which is very grammatical and conveys exactly the same meaning.

Algorithm If we had limited operations to just 1)-3), the minimum number of edits can be computed with dynamic programming (even when they have unequal costs) – **Levenstein** or **edit** distance. Adding the last operation however, makes the problem NP-hard, so an approximation is used.

Calculation phases are repeated for all references and the best (lowest) score is retained:

I min. number of insertions, deletions, and substitutions is calculated using dynamic programming.

II a greedy search is used to find the set of shifts, by repeatedly selecting the shift that most reduces the number of insertions, deletions and substitutions, until no more beneficial shifts remain.

III dynamic programming is reused to optimally calculate the remaining edit distance using a minimum-edit-distance over 3 basic operations

TER calculation algorithm:

Require: hypothesis h, references R

$E \leftarrow \infty$
for all $\forall r \in R$ **do**
  $h' \leftarrow h$
  $e \leftarrow 0$
  repeat
    {Find shift, s, that most reduces min-edit-distance(h', r)}
    if s reduces edit distance **then**
      $h' \leftarrow$ apply s to h
      $e \leftarrow e + 1$
    end if
  until no distance-reducing shifts remain
  $e \leftarrow e+$ min-edit-distance(h', r)
  if $e < E$ **then**
    $E \leftarrow e$
  end if
end for

In order to reduce the space of possible shifts (for efficiency), several constraints are used:

- The shifted words must match the words in the REF destination position exactly.
- The words of the HYP in the original position and the corresponding REF words must not exactly match.

- The words of the REF that correspond to the destination position must be misaligned before the shift (e.g., deleted or inserted, not substituted).

Example:

```
REF: a b c d e f   c
HYP: a     d e   b c f
```

The words b c in the hypothesis can be shifted to the left to correspond to the words b c in the reference, because there is a mismatch in the current location of b c in the hypothesis, and there is a mismatch of b c in the reference.

1.6 Statistical significance testing

Question: Are differences in BLEU for two systems random or not?

- Null hypothesis: The two systems are equal and observed difference is random.
- p-value: probability of incorrectly rejecting null hypothesis. A small p-value (≤ 0.05) means that observed difference is statistically significant, i.e., difference is not random.

Approximate Randomization Test:

Set $c = 0$
Compute actual statistic of score differences $|S_X - S_Y|$ on test data for system X, Y
for all random shuffles $r = 0, \ldots, R$ **do**
  for all sentences in test set **do**
    Shuffle variable tuples between system X and Y with probability 0.5
  end for
  Compute pseudo-statistic $|S_{X_r} - S_{Y_r}|$ on shuffled data
  if $|S_{X_r} - S_{Y_r}| \geq |S_X - S_Y|$ **then**
    c ++
  end if
end for
$p = (c + 1)/(R + 1)$
Reject null hypothesis if p is less than or equal to specified rejection level.

"Variable types" for BLEU are n-gram matches, n-gram counts and the length of candidate and reference translation for each candidate.
Idea: Under the null hypothesis, systems are not different, thus any variable tuple produced by one of the systems could have been produced just as well by the other system.
Significance levels p are computed by the percentage of trials where the test statistic on shuffled data is greater than the actual test statistic.

1.7 References

Callison-Burch, C., Osborne, M. and Koehn, P. (2006): *Re-evaluating the Role of Bleu in Machine Translation Research.* In: 11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006.

Riezler, S. and Maxwell, J. T. (2005): *On Some Pitfalls in Automatic Evaluation and Significance Testing for MT.* In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan.