MACHINE
TRANSLATION
EVALUATION

Advanced Signal Processing Seminar
Emine Zerrin SAKIR
Stefan Petrik
Overview

- Introduction to machine translation
- N-gram based methods
- BLUE
- NIST
- Word error rate based methods
- Minimum error rate training
A Brief Description of Machine Translation
Introduction

- Machine Translation (MT) is a subfield of computational linguistics.
- It investigates the use of computer software to translate text or speech from one natural language to another.
- The translation process, basically, includes two steps:
  1. Decoding the meaning of the source text
  2. Re-encoding this meaning in the target language
The Challenges of Machine Translation

- How to program a computer to understand a text as a human being does!
- To create a new text in the target language that sounds as if it has been written by a human being!
Approaches

- Lexicon-based machine translation
- Statistical machine translation
- Example-based machine translation
- Interlingual machine translation
interlingua

analysis

transfer

Direct translation

Source text

Target text

generation
Evaluation of MT Systems
Pros and Cons of Human Evaluation of Machine Translation

- Human evaluations of MT are extensive but expensive.
- Human evaluations of MT are too much time consuming which is not practical for developers.
- Human evaluations of MT take human labor which can not be reused.
- Human evaluations of MT weigh many aspects of translation: adequacy, fidelity, fluency
Some Methods of Automatic Evaluation of MT

- BLEU
- NIST
- METEOR
Descriptions

- **N-Gram**: It is a sub-sequence of n items from a given sequence.
- **Unigram**: n-gram of size 1.
- **Bigram**: n-gram of size 2.
- **Trigram**: n-gram of size 3.
BLEU

- BLEU: \textit{BiLingual Evaluation U}nderstudy
- The quality of translation is indicated as a number between 0 and 1.
- It is measured as statistical closeness to a given set of good quality human reference translations.
- It does not directly take into account translation intelligibility or grammatical correctness.
Viewpoint of „BLEU“ Method

- The criteria of translation performance measurement is:

  *The closer a machine translation is to a professional human translation, the better it is.*

- So, the MT evaluation system requires two ingredients:
  1. A numerical „translation closeness“ metric
  2. A corpus of good quality human reference translations
The Baseline BLEU Metric

Example 1:

- **Candidate 1**: It is a guide to action which ensures that the military always obeys the commands of the party.
- **Candidate 2**: It is to ensure the troops forever hearing the activity guidebook that party direct.
- **Reference 1**: It is a guide to action that ensures 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.
The Baseline BLEU Metric

- The primary programming task in BLEU implementation is:
  
  *To compare *n*-grams of the candidate with the *n*-grams of the reference translation and count the number of matches.*

- These matches are position independent.

- The more the matches, the better the candidate translation.
Modified Unigram Precision

- Example2:
  - Candidate: the the the the the the the the
  - Reference 1: The cat is on the mat.
  - Reference 2: There is a cat on the mat.
- The max. number of “the” is 2 in any single reference (Reference 2). So this number is clipped.
- Resulting modified unigram precision is: 2/7.
Modified n-gram Precision

- Modified n-gram precision computation for any n:
  - All candidate n-gram counts and their corresponding max. reference counts are collected.
  - The candidate counts are clipped by their corresponding reference max. value.
  - These values are summed and divided by the total number of candidate n-grams.
Modified n-gram Precision on Blocks of Text

- The modified n-gram precision on a multi-sentence test set is computed by the formula:

\[
p_n = \frac{\sum_{c \in \{\text{Candidates}\}} \sum_{n-\text{gram} \in C} \text{Countclip}(n-\text{gram})}{\sum_{c \in \{\text{Candidates}\}} \sum_{n-\text{gram} \in C} \text{Count}(n-\text{gram})}
\]

- This means that a word-weighted average of the sentence-level modified precision is used rather than a sentence-weighted average!
Ranking Systems Using Only Modified n-gram Precision

- The average modified precisions on the output of a human and machine translators.
- There are 4 reference translations for each of 127 source sentences.
Combining the n-gram Precisions

- As seen from the figure, the modified n-gram precision decays roughly, exponentially with n.
- BLEU uses the average logarithm with uniform weights, which is equivalent to using the geometric mean of the modified n-gram precisions.
Sentence Length

- A candidate translation length should not be too long or too short.
- Even though n-gram precision accomplishes this by penalizing using a word more times than it occurs in any of the reference, it alone fails to enforce the proper translation length.

Example 3:
- Candidate: of the
- Reference 1: It is a guide to action that ensures that the military will forever heed the 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.
The Trouble with Recall

- Reference translations may choose different words to translate the same source word and the candidate should not recall all the references.

- Example 4:
  - 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.
Sentence Brevity Penalty

- Brevity penalty factor penalizes candidates that are shorter than their reference.
- With this parameter in place, a high scoring candidate translation must match the reference translations in:
  - Length
  - Word choice
  - Word order
- Both n-gram precision length effect and brevity penalty considers the reference translation lengths in the target language.
Brevity Penalty

- Brevity penalty is a multiplicative factor, modifying the overall BLEU score.
- Brevity penalty is a decaying exponential in \( r/c \), where:
  - \( r \): test corpus’s effective reference length. It is computed by summing the best match lengths for each candidate sentence in the corpus.
  - \( c \): total length of the candidate translation corpus.
BLEU DETAILS

- The ranking behavior:
  \[ N=4, \ w_n=1/N \]

\[
BP = \begin{cases} 
1 & \text{if } c > r \\
\exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r 
\end{cases}
\]

\[
BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)
\]
The BLEU Evaluation

- The BLUE scores of the five systems against two references on the test corpus of 500 sentences.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>H1</td>
<td>H2</td>
</tr>
<tr>
<td>0.0527</td>
<td>0.0829</td>
<td>0.0930</td>
<td>0.1934</td>
<td>0.2571</td>
</tr>
</tbody>
</table>

- How reliable is the difference in BLUE metric?
- What is the variance of BLUE score?
- If another random set of 500 sentences were taken, would the results be same?
BLEU Evaluation

- The test corpus is divided into 20 blocks of 25 sentences and for each the BLEU metric is computed.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>H1</th>
<th>H2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.051</td>
<td>0.081</td>
<td>0.090</td>
<td>0.192</td>
<td>0.256</td>
</tr>
<tr>
<td>StdDev</td>
<td>0.017</td>
<td>0.025</td>
<td>0.020</td>
<td>0.030</td>
<td>0.039</td>
</tr>
<tr>
<td>t</td>
<td>-</td>
<td>6</td>
<td>3.4</td>
<td>24</td>
<td>11</td>
</tr>
</tbody>
</table>
NIST

- NIST is another method for evaluating the quality of the text translated using machine translation.
- NIST is based on BLEU metric with some alterations:
  - NIST calculates how informative a particular n-gram is.
  - When calculating brevity penalty small variations in translation length do not impact overall score very much.
The NIST Score Formulation

- Computation of information weights:

\[
Info(w_1...w_n) = \log_2 \left( \frac{\text{the number of occurrences of } w_1...w_n - 1}{\text{the number of occurrences of } w_1...w_n} \right)
\]

\[
Score = \sum_{n=1}^{N} \left\{ \sum_{\text{all } w_1...w_n \text{ that co-occur}} \frac{Info(w_1...w_n)}{(1)} \right\} \times \exp \left\{ \beta \log^2 \left[ \min \left( \frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\}
\]
Performance vs. Parameter Selection

- Performance as a function of source
- Performance vs. number of references
- Performance vs. segment size
- Performance with more language training
- Performance with preservation of case
- Performance with reference normalization
Conclusion

- The progress made in automatic evaluation of machine translation helps the developers.
- Provides MT a significant progress.
- Automatic machine translation evaluation can be developed for a more accurate estimator of translation based on current techniques.
References


Thanks for your attention…