International Workshop on Spoken Language Translation
Pittsburgh, USA
October 24-25, 2004

Evaluating Machine Translation Output with Automatic Sentence Segmentation

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University
D-52056 Aachen
1. related work
2. motivation
3. state-of-the-art translation error measures
4. document-level vs. sentence-level evaluation
5. description of the algorithm
6. experimental results
7. summary
Related work

- BLEU: (Papineni et al., 2001), NIST: (Doddington, 2002)
- preprocessing/normalization for MT Evaluation: (Leusch et al., 2005)
Motivation - 1

- evaluation plays a crucial role in machine translation research and acceptance of MT technology
- human evaluation is time-consuming and expensive
- automatic evaluation is preferred, but its quality still leaves to be desired
- some well-established evaluation measures exist
  - WER, PER, BLEU, NIST, ...
- all objective measures include the concept of sentences or segments
- all use (multiple) reference translations of these segments
- each evaluation algorithm expects exactly one target language segment for each source language segment
Motivation - 2

• concept of sentences is in general not well-defined for speech translation

• current situation in ASR+MT evaluations:
  – humans transcribe the acoustic signal and define segment boundaries
  – ASR systems are forced to generate segment boundaries at the same timeframes
  – segments may be too short/long, ASR/MT systems may lose context information

• more realistic conditions:
  – ASR system suggests segment boundaries based on prosodic or LM features
  – MT system may split or merge these segments to meet its constraints or modeling assumptions

• BUT: the segments in the produced translations and manual references will be different!

⇒ existing MT error measures will not be applicable
Solution

- align the output of an MT system with the reference translations
- re-segment the MT output based on the segmentation of the reference translations
- make use of the Levenshtein edit distance algorithm
- take multiple references into account
Existing Error Measures - 1

- All well-established measures are based on segment-level comparisons.
- Scores for the whole document are obtained by summation over all segments and normalization.
- \( W_{ER} \) is the Levenshtein (edit) distance.
- \( P_{ER} \) is similar to \( W_{ER} \), but ignores the order of words within a segment.
- \( W_{ER} \) and \( P_{ER} \) are normalized by the total reference length which can be computed in several ways (Leusch et al., 2005).
- \( B_{LEU} \) is an \( m \)-gram precision measure.
- \( N_{IST} \) extends \( B_{LEU} \) with information weights.
- \( B_{LEU} \) and \( N_{IST} \) use a global brevity penalty to avoid a bias towards short candidate translations.
Document-level vs. Segment-level Evaluation

- evaluation at document level:
  - assume the whole candidate document and each reference document to have only one segment
- computation of WER at document level is possible only using a single reference document (as in ASR)
- PER, BLEU and NIST can be computed at document level, but the estimates of translation quality will be too optimistic
  - e.g. an \( m \)-gram starting with the first word in the candidate document will match an \( m \)-gram starting with the 500th word in a reference document

⇒ segment-level evaluation is preferrable
  (with a proper definition of segments)
Alignment algorithm: Notation

- $w_1, \ldots, w_n, \ldots, w_N$ is a reference document segmented into $K$ segments
- reference segmentation is defined by indices $n_1, \ldots, n_k, \ldots, n_K := N$
- candidate document $e_1, \ldots, e_i, \ldots, e_I$
- goal: find a Levenshtein alignment between the two documents
- mark words which are aligned to $w_{n_k}$ and obtain the segmentation of the candidate document $i_1, \ldots, i_k, \ldots, i_K := I$. 
Dealing with Multiple References

- extend the algorithm to work with multiple reference documents $r = 1, \ldots, R$

- without loss of generality, assume that a reference translation of a segment $k$ has the same length across reference documents
  - achieved by inserting artificial “empty word” symbols $\$\$ at the end of reference segments which are shorter than the translation with the maximum length

- consequently, the reference words are indexed by $w_{nr}$, $n = 1, \ldots, N$, $r = 1, \ldots, R$
Algorithm: Within-segment Alignment

• for each candidate word index $i$, reference word index $n$, and reference index $r$, compute Levenshtein distance recursively with dynamic programming using the auxiliary quantity $D$:

$$D(i, n, r) = \min \begin{cases} 
D(i - 1, n - 1, r) + 1 - \delta(e_i, w_{nr}), \\
D(i - 1, n, r) + 1, \\
D(i, n - 1, r) + 1 - \delta(w_{nr}, \$) 
\end{cases}$$

• determine, which possibility has lower costs:
  – a match, substitution, insertion or deletion
• special case: deletion with no costs
  ⇒ reference that does not have the maximum length is already processed
• the index of the last locally best segment boundary is saved in a backpointer $B(i, n, r)$
• the backpointer of the best predecessor hypothesis is passed on in each recursion step
Recombination at Reference Segment Boundaries

- two consecutive candidate segments can be scored with segments from different reference documents

\[
D(i, n = n_k, r) = \min_{r' = 1, \ldots, R} D(i, n - 1, r')
\]

\[
BR(i, k) = \hat{r} = \arg\min_{r' = 1, \ldots, R} D(i, n - 1, r')
\]

\[
BP(i, k) = B(i, n - 1, \hat{r})
\]

- backpointers pass on the locally optimal reference and the hypothesized segment boundary for the segment \(k - 1\)
Automatic Segmentation Word Error Rate

- algorithm terminates by reaching the last word in candidate and each reference document
- the optimal number of edit operations is given by:

$$d_L = \min_r D(I, N, r)$$

- the sentence boundary decisions $i_1, \ldots, i_K$ and the optimal sequence of reference segments $\hat{r}_1, \ldots, \hat{r}_K$ are recursively backtraced using the backpointer arrays
- $\hat{r}_1, \ldots, \hat{r}_K$ is the new single-reference document $\hat{E}$ with length $\hat{N}$
- we define the automatic segmentation word error rate (AS-WER) by:

$$\text{AS-WER} = \frac{d_L}{\hat{N}}$$
Complexity

- memory: $O(N \cdot R + I \cdot K)$
- time: $O(N \cdot I \cdot R)$
- fast C++ implementation using integer word indices and costs
- e.g. 2-3 minutes and max. 400 MB of memory to align 20K words using two reference documents with 2643 segments (desktop PC)
Experimental results

- test the new evaluation method on the data from IWSLT 2004 and TC-STAR 2005 evaluations
- on the IWSLT 2004 data, compute correlation with human judgements
Evaluation Methodology

- currently, the “correct” manual segmentation of the candidate translation is available
- compute $\text{WER}$, $\text{PER}$, $\text{BLEU}$, $\text{NIST}$ using either manual or automatic segmentation
- compute correlation coefficients with human judgments
  - adequacy, fluency (Pearson’s $r$)
  - ranking (Kendall’s $\tau$)
- compare relative score changes
- compare absolute score values
## Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>TC-STAR</th>
<th>BTEC CE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source language</strong></td>
<td>Spanish</td>
<td>Chinese</td>
</tr>
<tr>
<td><strong>Target language</strong></td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td><strong>Segments</strong></td>
<td>2643</td>
<td>500</td>
</tr>
<tr>
<td><strong>Running words</strong></td>
<td>20164</td>
<td>3632</td>
</tr>
<tr>
<td><strong>Ref. translations</strong></td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td><strong>Avg. ref. length</strong></td>
<td>7.8</td>
<td>7.3</td>
</tr>
<tr>
<td><strong>Candidate systems</strong></td>
<td>4</td>
<td>20</td>
</tr>
</tbody>
</table>
• correlation is slightly better when automatic segmentation is used
Correlation with Fluency

- correlation is slightly better when automatic segmentation is used
\[ \Rightarrow \text{AS-measures are suitable for evaluation and ranking of MT systems} \]
## Error Measures on the TC-STAR task - 1

<table>
<thead>
<tr>
<th>Error measure:</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>WER [%]</td>
<td>37.4</td>
</tr>
<tr>
<td>AS-WER [%]</td>
<td>36.2</td>
</tr>
<tr>
<td>PER [%]</td>
<td>30.7</td>
</tr>
<tr>
<td>AS-PER [%]</td>
<td>30.6</td>
</tr>
<tr>
<td>BLEU [%]</td>
<td>51.1</td>
</tr>
<tr>
<td>AS-BLEU [%]</td>
<td>50.9</td>
</tr>
<tr>
<td>NIST</td>
<td>10.34</td>
</tr>
<tr>
<td>AS-NIST</td>
<td>10.29</td>
</tr>
<tr>
<td>Segmentation ER [%]</td>
<td>6.5</td>
</tr>
</tbody>
</table>

- automatic segmentation does not change the ranking of the four systems
- absolute values of AS-measures can be slightly lower/higher (e.g. depending on the normalization method)
- segmentation error rate is small and degrades only slightly with degrading WER
Error Measures on the TC-STAR task - 2

<table>
<thead>
<tr>
<th>Error measure</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>BLEU [%]</td>
<td>51.1</td>
</tr>
<tr>
<td>AS-BLEU [%]</td>
<td>50.9</td>
</tr>
<tr>
<td>BLEU at document level [%]</td>
<td>55.3</td>
</tr>
<tr>
<td>NIST</td>
<td>10.34</td>
</tr>
<tr>
<td>AS-NIST</td>
<td>10.29</td>
</tr>
<tr>
<td>NIST at document level</td>
<td>11.57</td>
</tr>
</tbody>
</table>

- **BLEU** and **NIST** scores on document level overestimate the performance of MT systems
  - moreover, the difference between systems is significantly underestimated
- **AS-BLEU** and **AS-NIST** give reliable estimates of translation quality
Conclusions

- A novel method of automatic sentence segmentation to be used for evaluation of MT quality
- MT output does not have to have the same segmentation as the reference translations
- Automatic segmentation is determined efficiently with a modified edit distance algorithm
- Multiple reference translations are taken into account
- Existing MT evaluation measures can be applied
- The measures computed using automatic segmentation correlate with human judgement at least as well as when manual segmentation is used
- The new evaluation method is especially important for evaluating translations of automatically recognized and segmented speech
- The method resolves the issue of different segmentation requirements of ASR and MT