Evaluating translation quality - part 2

Machine Translation
Lecture 10

Instructor: Chris Callison-Burch
TAs: Mitchell Stern, Justin Chiu

Website: mt-class.org/penn
Bilingual Evaluation Understudy

- Uses multiple reference translations
- Look for n-grams that occur anywhere in the sentence
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

2-gram precision = 10/17

| Hyp          | appeared calm when he was taken to the American plane , which will to Miami , Florida . |
n-gram precision

<table>
<thead>
<tr>
<th>Hyp</th>
<th>appeared calm when he was taken to the American plane, which will to Miami, Florida.</th>
</tr>
</thead>
</table>

1-gram precision = $\frac{15}{18} = 0.83$
2-gram precision = $\frac{10}{17} = 0.59$
3-gram precision = $\frac{5}{16} = 0.31$
4-gram precision = $\frac{3}{15} = 0.20$

- Geometric average

$$(0.83 \times 0.59 \times 0.31 \times 0.2)^{\frac{1}{4}} = 0.417$$

or equivalently

$$\exp(\ln 0.83 + \ln 0.59 + \ln 0.31 + \ln 0.2/4) = 0.417$$
<table>
<thead>
<tr>
<th>Ref 1</th>
<th>Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref 2</td>
<td>Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.</td>
</tr>
<tr>
<td>Ref 3</td>
<td>Orejuela appeared calm as he was being led to the American plane that was to carry him to Miami in Florida.</td>
</tr>
<tr>
<td>Ref 4</td>
<td>Orejuela seemed quite calm as he was being led to the American plane that would take him to Miami in Florida.</td>
</tr>
<tr>
<td>Hyp</td>
<td>to the American plane</td>
</tr>
<tr>
<td>Hyp</td>
<td>to the American plane</td>
</tr>
<tr>
<td>----------------------</td>
<td>------------------------</td>
</tr>
</tbody>
</table>

- 1-gram precision $= \frac{4}{4} = 1.0$
- 2-gram precision $= \frac{3}{3} = 1.0$
- 3-gram precision $= \frac{2}{2} = 1.0$
- 4-gram precision $= \frac{1}{1} = 1.0$

$$\exp(\ln 1 + \ln 1 + \ln 1 + \ln 1) = 1$$
Brevity Penalty

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

- \(c\) is the length of the corpus of hypothesis translations.
- \(r\) is the effective reference corpus length.
- The effective reference corpus length is the sum of the single reference translation from each set that is closest to the hypothesis translation.
Brevity Penalty

BP

MT is Longer

MT is Shorter

Difference with effective reference length (%)
<table>
<thead>
<tr>
<th>Ref 1</th>
<th>Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.</th>
<th>r = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyp</td>
<td>appeared calm when he was taken to the American plane, which will go to Miami, Florida.</td>
<td>c = 18</td>
</tr>
<tr>
<td></td>
<td>$BP = \exp(1-(20/18)) = 0.89$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ref 1</th>
<th>Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.</th>
<th>r = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyp</td>
<td>to the American plane</td>
<td>c = 4</td>
</tr>
<tr>
<td></td>
<td>$BP = \exp(1-(20/4)) = 0.02$</td>
<td></td>
</tr>
</tbody>
</table>
Bleu = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)

• Geometric average of the n-gram precisions
• Optionally weight them with w
• Multiplied by the brevity penalty
### BLEU

<table>
<thead>
<tr>
<th>Hyp</th>
<th>appeared calm when he was taken to the American plane, which will to Miami, Florida.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \exp(1-(20/18)) \times \exp((\ln 0.83 + \ln 0.59 + \ln 0.31 + \ln 0.2)/4) = 0.374 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hyp</th>
<th>to the American plane</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \exp(1-(20/4)) \times \exp((\ln 1 + \ln 1 + \ln 1 + \ln 1)/4) = 0.018 )</td>
</tr>
</tbody>
</table>
Problems with BLEU

• (Discuss with your neighbor)
Problems with BLEU

- Synonyms and paraphrases are only handled if they are in the set of multiple reference translations.
- The scores for words are equally weighted so missing out on content-bearing material brings no additional penalty.
- The brevity penalty is a stop-gap measure to compensate for the fairly serious problem of not being able to calculate recall.
More Metrics

- WER - word error rate
- PI-WER - position independent WER
- METEOR - Metric for Evaluation of Translation with Explicit ORdering
- TERp - Translation Edit Rate plus
Even More Metrics

<table>
<thead>
<tr>
<th>Metric IDs</th>
<th>Participant</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER, AMBER-NL, AMBER-IT</td>
<td>National Research Council Canada (Chen and Kuhn, 2011)</td>
</tr>
<tr>
<td>F15, F15G3</td>
<td>Koç University (Bicici and Yuret, 2011)</td>
</tr>
<tr>
<td>METEOR-1.3-ADQ, METEOR-1.3-RANK</td>
<td>Carnegie Mellon University (Denkowski and Lavie, 2011a)</td>
</tr>
<tr>
<td>MTeRater, MTeRater-Plus</td>
<td>Columbia / ETS (Parton et al., 2011)</td>
</tr>
<tr>
<td>mp4IBM1, mpF, wmpF</td>
<td>DFKI (Popović, 2011; Popović et al., 2011)</td>
</tr>
<tr>
<td>PARSECONF</td>
<td>DFKI (Avramidis et al., 2011)</td>
</tr>
<tr>
<td>ROSE, ROSE-POS</td>
<td>The University of Sheffield (Song and Cohn, 2011)</td>
</tr>
<tr>
<td>TESLA-B, TESLA-F, TESLA-M</td>
<td>National University of Singapore (Dahlmeier et al., 2011)</td>
</tr>
<tr>
<td>TINE</td>
<td>University of Wolverhampton (Rios et al., 2011)</td>
</tr>
<tr>
<td>BLEU</td>
<td>provided baseline (Papineni et al., 2002)</td>
</tr>
<tr>
<td>TER</td>
<td>provided baseline (Snover et al., 2006)</td>
</tr>
</tbody>
</table>
How do we know which metric is best?

- Measure correlation with human judgments
- How do people evaluation MT quality
## Manual Evaluation

**Source:** Estos tejidos están analizados, transformados y congelados antes de ser almacenados en Hema-Québec, que gestiona también el único banco público de sangre del cordón umbilical en Quebec.

**Reference:** These tissues are analyzed, processed and frozen before being stored at Héma-Québec, which manages also the only bank of placental blood in Quebec.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>These weavings are analyzed, transformed and frozen before being stored in Hema-Quebec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.</td>
<td>![Rank](1 2 3 4 5)</td>
</tr>
<tr>
<td>These tissues analysed, processed and before frozen of stored in Hema-Québec, which also operates the only public bank umbilical cord blood in Quebec.</td>
<td>![Rank](1 2 3 4 5)</td>
</tr>
<tr>
<td>These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also manages the only public bank umbilical cord blood in Quebec.</td>
<td>![Rank](1 2 3 4 5)</td>
</tr>
<tr>
<td>These tissues are analyzed, processed and frozen before being stored in Hema-Québec, which also operates the only public bank of umbilical cord blood in Quebec.</td>
<td>![Rank](1 2 3 4 5)</td>
</tr>
<tr>
<td>These fabrics are analyzed, are transformed and are frozen before being stored in Hema-Québec, who manages also the only public bank of blood of the umbilical cord in Quebec.</td>
<td>![Rank](1 2 3 4 5)</td>
</tr>
</tbody>
</table>
5-point scales

**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

Table 3: The scales for manually assigned adequacy and fluency scores

Table 4: Two hypothesis translations with similar Bleu scores but different human scores, and one of four reference translations

The manual evaluation conducted for the NIST MT Eval is done by English speakers without reference to the original Arabic or Chinese documents. Two judges assigned each sentence in

The output of the system that was ranked 1st by Bleu is not publicly available.
Heather Locklear Arrested for driving under the influence of drugs

The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses."

Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw Ms. Locklear's life.

- Why was Heather Locklear arrested?
  - She was arrested on suspicion of driving under the influence of drugs.
- Why did the bystander call emergency services?
  - He was concerned for Ms. Locklear’s life.
- Where did the witness see her acting abnormally?
  - Pulling out of parking in Montecito
Heather Locklear, known for her role in the TV series "Melrose Place," was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness observed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesperson for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses."

Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw Ms. Locklear 'pressed after 16:30 clock accelerator and a lot of noise did when she attempted to move their car towards behind or forward from the parking space, and when it went backwards, she pulled itself together unites Male at their sunglasses'. A little later the female witness that did probably have known the actress saw her act abnormally in a parking lot.
Heather Locklear Arrested for Driving Under the Influence of Drugs

The actress Heather Locklear, known for her role in the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4:30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses."

Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw Ms. Locklear driving under the effect of an unknown medicine. The actress, known for the role of Amanda from the series "Melrose Place," was arrested at this weekend in Santa Barbara (California) because of driving under the effect of an unknown medicine. A witness had observed how she attempted quite strange way how to go from their parking space in Montecito, reported spokesman for the traffic police of California to the magazine 'People'. The witness told in detail that Locklear 'after 16:30 clock through pedal and a lot of noise did when she attempted to move their car towards behind or forward from the parking space, and when she went backwards, took it a few times in their sunglasses'.

The traffic police of California told People magazine about the details of the witness's story and the actress's arrest. The witness probably recognized that Locklear on a nearby road and stopped the car with the witness to the car off.
Heather Locklear Arrested for driving under the influence of drugs

The actress Heather Locklear, Amanda of the popular series Melrose Place, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness viewed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear “hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses.” Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw...
Heather Locklear Arrested for driving under the influence of drugs

The actress Heather Locklear, known through the popularity of the series "Melrose Place" as Amanda, was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness observed her performing inappropriate maneuvers while trying to take her car out from a parking in Montecito, as revealed to People magazine by a spokesman for the Californian Highway Police. The witness stated that around 4.30pm Ms. Locklear "hit the accelerator very violently, making excessive noise while trying to take her car out from the parking with abrupt back and forth maneuvers. While reversing, she passed several times in front of his sunglasses."

Shortly after, the witness, who, in a first time, apparently had not recognized the actress, saw Ms. Locklear using an unknown drug, arrested by the Californian police. The witness told in detail that Locklear "after 16:30 clock pressed 'accelerator and a lot of noise did when she attempted to move their car towards behind or forward from the parking space, and when it went backwards, took it a few times in their Sunglass'. A little later the witness saw the actress probably had not recognized that Locklear on a nearby road and stopped the car."

Medikamentes unknown have the effect of a fahrens under actress heather locklear arrested in Santa. One is, melrose place the series of the role of the 'remember the locklear actress the heather this weekend, because of the fahrens Barbara (California) in effect unknown medikamentes arrested People 'magazine. The traffic police California, spokesman for the auszufahren montecito reported in its way from tried parklücke type strange right, you have seen as a witness. . In some Zeitung, as and when they tried to a great deal of 30 p.m., witness the detail of history locklear after 16: that durchdrückte peddle noise and its progress was made parklücke for the car or moving backwards, they had they times of their sonnenbrille ' . The first was probably recognised that locklear a nearby road and anhielt, had not, with the witness to the car off.

Why was Heather Locklear arrested?

- She was arrested on suspicion of driving under the influence of drugs.
- Driving under the influence
- Driving while medicated
- DUI
- Driving while using drugs

Correct Answers

<table>
<thead>
<tr>
<th>System</th>
<th>Correct Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>94%</td>
</tr>
<tr>
<td>Google</td>
<td>80%</td>
</tr>
<tr>
<td>RBMT5</td>
<td>77%</td>
</tr>
<tr>
<td>Geneva</td>
<td>63%</td>
</tr>
<tr>
<td>JHU - Tromble</td>
<td>50%</td>
</tr>
</tbody>
</table>
The man was on assignment from the Ministry of Defense when he left two highly classified documents on a train to Waterloo.

The man was seconded by the Ministry of Defense when he was two extremely confidential documents in a train to Waterloo lost.

The man was working for the Ministry of Defense when he lost two extremely confidential documents in a train to Waterloo.
Reading Comprehension of Machine Translation

- Jones et al (2005) - Measured translation quality by testing English speakers on a Defense Language Proficiency Test for Arabic
- Read the MT output, and assess how many questions were answered correctly
- Nice, intuitive gauge of how good MT quality actually is
Which type of Human Evaluation is Best?

The agreement on the other two types of manual evaluation that we introduced were considerably better. The both the sentence and constituent ranking had moderate inter-annotator agreement and substantial intra-annotator agreement. Because the constituent ranking examined the translations of short phrases, often times all systems produced the same translations. Since these trivially increased agreement (since they would always be equally ranked) we also evaluated the inter- and intra-annotator agreement when those items were excluded. The agreement remained very high for constituent-based evaluation.

6.2 Timing

We used the web interface to collect timing information. The server recorded the time when a set of sentences was given to a judge and the time when the judge returned the sentences. We divided the time that it took to do a set by the number of sentences in the set. The average amount of time that it took to assign fluency and adequacy to a single sentence was 26 seconds. The average amount of time it took to rank a sentence in a set was 20 seconds. The average amount of time it took to rank a highlighted constituent was 11 seconds. Figure 4 shows the distribution of times for these tasks.

6.3 Correlation between automatic metrics and human judgments

To measure the correlation of the automatic metrics with the human judgments of translation quality we used Spearman's rank correlation coefficient. We opted for Spearman rather than Pearson because it makes fewer assumptions about the data. Importantly, it can be applied to ordinal data (such as the fluency and adequacy scales). Spearman's rank correlation coefficient is equivalent to Pearson correlation on ranks.

After the raw scores that were assigned to systems by an automatic metric and by one of our manual evaluation techniques have been converted to ranks, we can calculate using the simplified equation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

where $d_i$ is the difference between the rank for system $i$ and $n$ is the number of systems. The possible values of $\rho$ range between 1 (where all systems are ranked in the same order) and $-1$ (where the systems are ranked in the reverse order). Thus an automatic evaluation metric with a higher value for $\rho$ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower $\rho$.

Table 17 reports for the metrics which were used to evaluate translations into English.

Table 7 summarizes the results by averaging the correlation numbers by equally weighting each of the data conditions. The table ranks the automatic evaluation metrics based on how well they correlated with human judgments. While these are based on a relatively few number of items, and while we have not performed any tests to determine whether the differences in $\rho$ are statistically significant, the results.

The Czech-English conditions were excluded since there were so few systems.
Using manual judgments to evaluate automatic metrics...

- Measure correlation with human judgments
- System-level correlation
- Sentence-level correlation
Calculating Correlation

- The human evaluation metrics provide a ranking of the systems
  - So do the automatic metrics
- Calculate the correlation between the two lists
  - Metrics with higher correlation better predict human judgments
Spearman’s rank correlation coefficient

• For system-level correlation

\[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]
Kendall’s Tau

- Segment level evaluation

\[ \tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \]

Table 14: Segment-level Kendall’s tau correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average correlation.

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>METEOR-1.3-RANK</th>
<th>METEOR-1.3-ADQ</th>
<th>MTE _RATER-PL _LUS</th>
<th>TESLA-_F</th>
<th>TESLA-_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN-_FR</td>
<td>0.23</td>
<td>0.25</td>
<td>0.38</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>EN-_DE</td>
<td>0.24</td>
<td>0.25</td>
<td>0.38</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>EN-_ES</td>
<td>0.24</td>
<td>0.25</td>
<td>0.38</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>EN-_CZ</td>
<td>0.24</td>
<td>0.25</td>
<td>0.38</td>
<td>0.28</td>
<td>0.24</td>
</tr>
</tbody>
</table>

Table 15: Segment-level Kendall’s tau correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average correlation.

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>METEOR-1.3-RANK</th>
<th>METEOR-1.3-ADQ</th>
<th>MTE _RATER-PL _LUS</th>
<th>TESLA-_F</th>
<th>TESLA-_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER-_TI</td>
<td>0.32</td>
<td>0.22</td>
<td>0.31</td>
<td>0.21</td>
<td>0.27</td>
</tr>
<tr>
<td>AMBER</td>
<td>0.31</td>
<td>0.21</td>
<td>0.31</td>
<td>0.22</td>
<td>0.26</td>
</tr>
<tr>
<td>MP-F</td>
<td>0.31</td>
<td>0.22</td>
<td>0.30</td>
<td>0.20</td>
<td>0.26</td>
</tr>
<tr>
<td>WMP-F</td>
<td>0.31</td>
<td>0.22</td>
<td>0.29</td>
<td>0.19</td>
<td>0.25</td>
</tr>
<tr>
<td>AMBER-_NL</td>
<td>0.30</td>
<td>0.19</td>
<td>0.29</td>
<td>0.20</td>
<td>0.25</td>
</tr>
<tr>
<td>METEOR-_RANK</td>
<td>0.31</td>
<td>0.14</td>
<td>0.26</td>
<td>0.19</td>
<td>0.23</td>
</tr>
<tr>
<td>F15-_G</td>
<td>0.26</td>
<td>0.08</td>
<td>0.22</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>F15</td>
<td>0.26</td>
<td>0.07</td>
<td>0.22</td>
<td>0.12</td>
<td>0.17</td>
</tr>
<tr>
<td>MP4</td>
<td>0.25</td>
<td>0.12</td>
<td>0.13</td>
<td>0.06</td>
<td>0.15</td>
</tr>
<tr>
<td>IBM-1</td>
<td>0.21</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>0.15</td>
</tr>
<tr>
<td>DFKI-_PARSECONF</td>
<td>n/a</td>
<td>0.24</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

The metrics that had the strongest correlation this year included two metrics, MTeRater and TINE, as well as metrics that have demonstrated strong correlation in previous years like TESLA and Meteor.

6.2 Segment-Level Metric Analysis

We measured the metrics' segment-level scores with the human rankings using Kendall’s tau rank correlation coefficient. The reference was not included as an extra translation.

We calculated Kendall’s tau as:

\[ \tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \]

where a concordant pair is a pair of two translations of the same segment in which the ranks calculated from the same human ranking task and from the corresponding metric scores agree; in a discordant pair, they disagree.

In order to account for accuracy- vs. error-based metrics correctly, counts of concordant vs. discordant pairs were calculated specific to these two metric types. The possible values of \( \tau \) range between 1 (where all pairs are concordant) and -1 (where all pairs are discordant). Thus an automatic evaluation metric with a higher value for \( \tau \) is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower value.

We did not include cases where the human ranking was tied for two systems. As the metrics produce absolute scores, compared to five relative ranks in the human assessment, it would be potentially unfair to the metric to count a slightly different metric score as discordant with a tie in the relative human rankings. A tie in automatic metric rank for two translations was counted as discordant with two corresponding non-tied human judgments.

The correlations are shown in Table 14 for translations into English, and Table 15 out of English, sorted by average correlation across the four language pairs. The highest correlation for each language pair and the highest overall average are shown.
Many metrics are better than BLEU
Related Work

A number of projects in the past have looked into ways of extending and improving the Bleu metric. Doddington (2002) suggested changing Bleu’s weighted geometric average of n-gram matches to an arithmetic average, and calculating the brevity penalty in a slightly different manner. Hovy and Ravichandra (2003) suggested increasing Bleu’s sensitivity to inappropriate phrase movement by matching part-of-speech tag sequences against reference translations in addition to Bleu’s n-gram matches. Babych and Hartley (2004) extend Bleu by adding frequency weighting to lexical items through TF/IDF as a way of placing greater emphasis on content-bearing words and phrases.

Two alternative automatic translation evaluation metrics do a much better job at incorporating recall than Bleu does. Melamed et al. (2003) formulate a metric which measures translation accuracy in terms of precision and recall directly rather than precision and a brevity penalty. Banerjee and Lavie (2005) introduce the Meteor metric, which also incorporates recall on the unigram level and further provides facilities incorporating stemming, and WordNet synonyms as a more flexible match.

Lin and Hovy (2003) as well as Soricut and Brill (2004) present ways of extending the notion of n-gram co-occurrence statistics over multiple references, such as those used in Bleu, to other natural language generation tasks such as summarization. Both these approaches potentially suffer from the same weaknesses that Bleu has in machine translation evaluation.

Coughlin (2003) performs a large-scale investigation of Bleu’s correlation with human judgments, and finds one example that fails to correlate. Her future work section suggests that she has preliminary evidence that statistical machine translation systems receive a higher Bleu score than their non-n-gram-based counterparts.

Conclusions

In this paper we have shown theoretical and practical evidence that Bleu may not correlate with human judgment to the degree that it is currently believed to do. We have shown that Bleu’s rather coarse model of allowable variation in translation can mean that an improved Bleu score is not sufficient to reflect a genuine improvement in translation quality. We have further shown that it is not necessary to receive a higher Bleu score in order to be judged to have better translation quality by human subjects, as illustrated in the 2005 NIST Machine Translation Evaluation and our experiment manually evaluating Systran and SMT translations.

What conclusions can we draw from this? Should we give up on using Bleu entirely? We think that the advantages of Bleu are still very strong; automatic evaluation metrics are inexpensive, and do allow many tasks to be performed that would otherwise be impossible. The important thing therefore is to recognize which uses of Bleu are appropriate and which uses are not. Appropriate uses for Bleu include tracking broad, incremental changes to a single system, comparing systems which employ similar translation strategies (such as comparing phrase-based statistical machine translation systems with other phrase-based statistical machine translation systems), and using Bleu as an objective function to optimize the values of parameters such as feature weights in log linear translation models, until a better metric has been proposed.

Inappropriate uses for Bleu include comparing systems which employ radically different strategies (especially comparing phrase-based statistical machine translation systems against systems that do not employ similar n-gram-based approaches), trying to detect improvements for aspects of translation that are not modeled well by Bleu, and monitoring improvements that occur infrequently within a test corpus.

These comments do not apply solely to Bleu.
Re-evaluating the Role of BLEU in Machine Translation Research

Chris Callison-Burch  Miles Osborne  Philipp Koehn

If Bleu’s correlation with human judgments has been overestimated, then the field needs to ask itself whether it should continue to be driven by Bleu to the extent that it currently is. In this paper we give a number of counterexamples for Bleu’s correlation with human judgments. We show that under some circumstances an improvement in Bleu is not sufficient to reflect a genuine improvement in translation quality, and in other circumstances that it is not necessary to improve Bleu in order to achieve a noticeable improvement in translation quality.
Final thoughts on Evaluation
When writing a paper

- If you're writing a paper that claims that
  - one approach to machine translation is better than another, or that
  - some modification you've made to a system has improved translation quality

- Then you need to back up that claim

- Evaluation metrics can help, but good experimental design is also critical
Experimental Design

• Importance of separating out training / test / development sets
• Importance of standardized data sets
• Importance of standardized evaluation metric
• Error analysis
• Statistical significance tests for differences between systems
Evaluation drives MT research

- Metrics can drive the research for the topics that they evaluate
- NIST MT Eval -> DARPA Funding
- Bleu has lead to a focus on phrase-based translation
- Minimum error rate training (next lecture!)
- Other metrics may similarly change the community's focus
Invent your own evaluation metric

- If you think that Bleu is inadequate then invent your own automatic evaluation metric
- Can it be applied automatically?
- Does it correlate better with human judgment?
- Does it give a finer grained analysis of mistakes?
Goals for Automatic Evaluation

- No cost evaluation for incremental changes
- Ability to rank systems
- Ability to identify which sentences we're doing poorly on, and categorize errors
- Correlation with human judgments
- Interpretability of the score
- Quick to calculate for MERT
Questions?

- Tons of data available at
  - [http://statmt.org/wmt10/results.html](http://statmt.org/wmt10/results.html)
  - [http://statmt.org/wmt12/results.html](http://statmt.org/wmt12/results.html)
  - [http://statmt.org/wmt13/results.html](http://statmt.org/wmt13/results.html)
• Read 8 from the textbook