Evaluating Translation Quality

February 23, 2012
Goals for this lecture

• Understanding advantages of human versus automatic evaluation
• Details of BLEU
• How to validate automatic evaluation metrics
• What makes a good {manual / automatic} evaluation?
Evaluating MT Quality

• Why do we want to do it?
  ‣ Want to rank systems
  ‣ Want to evaluate incremental changes
  ‣ What to make scientific claims

• How not to do it
  ‣ “Back translation”
  ‣ The vodka is not good
Human Evaluation of MT v. Automatic Evaluation

- Human evaluation is
  - Ultimately what we're interested in, but
  - Very time consuming
  - Not re-usable

- Automatic evaluation is
  - Cheap and reusable, but
  - Not necessarily reliable
**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-Québec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.</td>
<td>![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>![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>![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>![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>![1 2 3 4 5]</td>
</tr>
</tbody>
</table>
### 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-Québec, that negotiates also the public only bank of blood of the umbilical cord in Quebec.</td>
<td><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></td>
</tr>
</tbody>
</table>
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
Methodology

• Comparison against reference translations
• Intuition: closer we get to human translations, the better we're doing
• Could use WER like in speech recognition
Word Error Rate

- Levenshtein Distance (also "edit distance")
- Minimum number of insertions, substitutions, and deletions needed to transform one string into another
- Useful measure in speech recognition
  - This shows how easy it is to recognize speech
  - This shows how easy it is to wreck a nice beach
Problems with using WER for translation?

- (discuss with your neighbor)
Problems with WER

- Unlike speech recognition we don't have the assumption of
  - exact match against the reference
- In machine translation there can be many possible (and equally valid) ways of translating a sentence
  - This shows how easy it is to recognize speech
  - It illustrates how simple it is to transcribe the spoken word
Problems with WER

- Unlike speech recognition we don't have the assumption of **linearity**
- Clauses can move around, since we're not doing transcription
  - This shows how easy it is to recognize speech
  - It is easy to recognize speech, as this shows
  - This shows that recognizing speech is easy
Solutions?

• (Talk to your neighbor)
Solutions

- Compare against lots of test sentences
- Use multiple reference translations for each test sentence
- Look for phrase / n-gram matches, allow movement
• **BiLingual Evaluation Understudy**
• Uses multiple reference translations
• Look for n-grams that occur anywhere in the sentence
### Multiple references

<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>
</tbody>
</table>
n-gram precision

\[
p_n = \frac{\sum_{S \in C} \sum_{ngram \in S} Count_{matched}(ngram)}{\sum_{S \in C} \sum_{ngram \in S} Count(ngram)}
\]

- BLEU modifies this precision to eliminate repetitions that occur across sentences.
<table>
<thead>
<tr>
<th>Ref</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.</td>
</tr>
<tr>
<td>2</td>
<td>Orejuela appeared calm while being escorted to the plane that would take him to Miami, Florida.</td>
</tr>
<tr>
<td>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>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>
</tbody>
</table>

“to Miami” can only be counted as correct once.
<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>appeared calm when he was taken to the American plane, which will to Miami, Florida.</td>
</tr>
</tbody>
</table>
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, will, would, ,, .

Hyp appeared calm when he was taken to the American plane, which will to Miami, Florida.
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>

<table>
<thead>
<tr>
<th>n-gram precision</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram precision</td>
<td>15/18 = .83</td>
</tr>
<tr>
<td>2-gram precision</td>
<td>10/17 = .59</td>
</tr>
<tr>
<td>3-gram precision</td>
<td>5/16 = .31</td>
</tr>
<tr>
<td>4-gram precision</td>
<td>3/15 = .20</td>
</tr>
</tbody>
</table>

- Geometric average
  \[
  \exp(\log .83 + \log .59 + \log .31 + \log .2) = 0.22
  \]
<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>
</tbody>
</table>
Better?

<table>
<thead>
<tr>
<th>Hyp</th>
<th>to the American plane</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram precision = ( \frac{4}{4} = 1.0 )</td>
<td></td>
</tr>
<tr>
<td>2-gram precision = ( \frac{3}{3} = 1.0 )</td>
<td></td>
</tr>
<tr>
<td>3-gram precision = ( \frac{2}{2} = 1.0 )</td>
<td></td>
</tr>
<tr>
<td>4-gram precision = ( \frac{1}{1} = 1.0 )</td>
<td></td>
</tr>
</tbody>
</table>

\[
\exp(\log 1 + \log 1 + \log 1 + \log 1) = 1
\]
Brevity Penalty

\[ BP = \begin{cases} 
1 & \text{if } c > r \\
\frac{e^{1-r/c}}{c} & \text{if } c \leq r
\end{cases} \]

- \( c \) is the length of the corpus of hypothesis translations
- \( r \) is the effective reference corpus length
Brevity Penalty

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

- \(c\) 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

MT is Longer

MT is Shorter

Difference with effective reference length (%)
<table>
<thead>
<tr>
<th>Ref</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>Hyp</td>
<td>appeared calm when he was taken to the American plane, which will take him to Miami, Florida.</td>
</tr>
</tbody>
</table>

\[
BP = \exp(1 - \frac{20}{18}) = 0.89
\]

<table>
<thead>
<tr>
<th>Ref</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>Hyp</td>
<td>to the American plane</td>
</tr>
</tbody>
</table>

\[
BP = \exp(1 - \frac{20}{4}) = 0.02
\]
Bleu = BP * exp(\(\sum_{n=1}^{N} w_n \log p_n\))

- Geometric average of the n-gram precisions
- Optionally weight them with w
- Multiplied by the brevity penalty
<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>
|     | \[ \exp(1-(20/18)) \times \exp(\log .83 + \log .59 + \log .31 + \log .2)\]
|     | \[ = 0.197 \]                                                                     |

<table>
<thead>
<tr>
<th>Hyp</th>
<th>to the American plane</th>
</tr>
</thead>
</table>
|     | \[ \exp(1-(20/4)) \times \exp(\log 1 + \log 1 + \log 1 + \log 1)\]
|     | \[ = 0.018 \]                                                                     |
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>![Circle] ![Circle] ![Circle] ![Circle] ![Circle] 5</td>
</tr>
</tbody>
</table>
**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><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></td>
</tr>
<tr>
<td>These tissues are analyzed, processed and frozen before being stored in Hema-Quebec, which also operates the only public bank of umbilical cord blood in Quebec.</td>
<td><img src="#" alt="Rank" /></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><img src="#" alt="Rank" /></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
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
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
Why was Heather Locklear arrested?

She was arrested on suspicion of 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
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

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.
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...
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...
The actress Heather Locklear that is known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (Californium) because of the driving under the effect of an unknown medicine. A female witness observed she attempted in quite strange way how to go from their parking space in Montecito, speaker of the traffic police of californium told the warehouse `People'. The female witness told in detail, that 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
The actress Heather Locklear that is known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (Californium) because of the driving under the effect of an unknown medicine. A female witness observed she attempted in quite strange way how to go from their parking space in Montecito, speaker of the traffic police of californium told the warehouse `People'. The female witness told in detail, that 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

- Why was Heather Locklear arrested?
  Driving while medicated

- Why did the bystander call emergency services?
  There was a lot of noise

- Where did the witness see her acting abnormally?
  In a parking lot
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

Was arrested actress Heather Locklear because of the driving under the effect of an unknown medicine

The actress Heather Locklear that is known to the Amanda through the role from the series "Melrose Place" was arrested at this weekend in Santa Barbara (Californium) because of the driving under the effect of an unknown medicine. A female witness observed she attempted in quite strange way how to go from their parking space in Montecito, speaker of the traffic police of californium told the warehouse ‘People’. The female witness told in detail, that 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

• Why was Heather Locklear arrested?
• Why did the bystander call emergency services?
• Where did the witness see her acting abnormally?
Actress Heather Locklear was due to driving under the influence of an unknown drug arrested

Actress Heather Locklear, by the role of Amanda from the series "Melrose Place" is known, was this weekend in Santa Barbara (California) because of driving under the influence of an unknown drug arrested. A witness had observed how it quite strange way tried to park their extended gap in Montecito, reported spokesman for the traffic police from California to the magazine 'People'. The witness told in detail that Locklear 'after 16:30 clock durchdrückte pedal and a lot of noise made by trying to her car to the rear or front of the park gap to move, and when she went backwards, took it a few times in their Sunglass'. Somewhat later the witness saw the beginning of the actress

- Why was Heather Locklear arrested?
- Why did the bystander call emergency services?
- Where did the witness see her acting abnormally?
Heather Locklear Arrested for Driving Under the Influence of Drugs

The actress Heather Locklear, known through the role of Amanda from the 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 4:30 pm, 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, was arrested actress Heather Locklear because of the driving under the effect of an unknown medicine.

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

---

1. Why was Heather Locklear arrested?
2. Why did the bystander call emergency services?
3. Where did the witness see her acting abnormally?
Heather Locklear Arrested for Driving Under the Influence of Drugs

The actress Heather Locklear, known for her role on 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...
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
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 impression, noticed that 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, took it a few times in their Sunglass'.

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

Medikamentes
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 account, was probably recognized as an actress, Locklear "after 16:30 clock durchdrückte pedal and a lot of noise made by trying to her car to the rear or front of the parking space, and when it went backwards, took it a few times in their Sunglass.' In some time, the witness saw the beginning of the actress. 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.
Heather Locklear, known for her role as Amanda in the popular series "Melrose Place," was arrested this weekend in Santa Barbara (California) after driving under the influence of drugs. A witness watched her perform 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...

Why was Heather Locklear arrested?

She was arrested on suspicion of driving under the influence of drugs.
Reference translation

The man was on assignment from the Ministry of Defense when he left two highly classified documents on a train to Waterloo.

Machine translation

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

Edited machine translation

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 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.
Which type of Human Evaluation is Best?

<table>
<thead>
<tr>
<th>Evaluation type</th>
<th>$P(A)$</th>
<th>$P(E)$</th>
<th>$K$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency (absolute)</td>
<td>.400</td>
<td>.2</td>
<td>.250</td>
</tr>
<tr>
<td>Adequacy (absolute)</td>
<td>.380</td>
<td>.2</td>
<td>.226</td>
</tr>
<tr>
<td>Fluency (relative)</td>
<td>.520</td>
<td>.333</td>
<td>.281</td>
</tr>
<tr>
<td>Adequacy (relative)</td>
<td>.538</td>
<td>.333</td>
<td>.307</td>
</tr>
<tr>
<td>Sentence ranking</td>
<td>.582</td>
<td>.333</td>
<td>.373</td>
</tr>
<tr>
<td>Constituent ranking</td>
<td>.693</td>
<td>.333</td>
<td>.540</td>
</tr>
</tbody>
</table>
Which type of Human Evaluation is Best?

![Graph showing distributions of time taken to judge sentences](image)

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.

Sets which took longer than 5 minutes were excluded from these calculations, because there was a strong chance that annotators were interrupted while completing the task.

These timing figures are promising because they indicate that the tasks which the annotators were the most reliable on (constituent ranking and sentence ranking) were also much quicker to complete than the ones that they were unreliable on (assigning fluency and adequacy scores). This suggests that fluency and adequacy should be replaced with ranking tasks in future evaluation exercises.

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:

\[ r_s = 1 \left( \frac{\sum (d_i^2)}{n(n^2-1)} \right) \]

where \( d_i \) is the difference between the rank for system \( i \) and \( n \) is the number of systems. The possible values of \( r_s \) 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 \( r_s \) is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower \( r_s \).

Table 17 reports \( r_s \) for the metrics which were used to evaluate translations into English.
Back to 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

\[ \tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \]
Many metrics are better than BLEU.
This is bad

![Graph showing human score against Bleu score for different systems.](image)

**Figure 4:** Bleu scores plotted against human judgments of fluency and adequacy, showing that Bleu vastly underestimates the quality of a non-statistical system.

### 5 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.

### 6 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