Natural Language Processing: Machine Translation

Borrows some slides from Kevin Knight, Dan Klein, and Bill MacCartney
Lecture Plan

1. The IBM (Alignment) Models [30 mins]
2. Middle 10 mins: Administration, questions, catch up [10 mins]
3. Getting parallel sentences to train on [10 mins]
4. Searching for the best translation: Decoding [10 mins]
5. MT Evaluation [10 mins]
1. The IBM (Alignment) Models [30 mins]
2. Middle 10 mins: Administration, questions, catch up [10 mins]
3. Getting parallel sentences to train on [10 mins]
4. Searching for the best translation: Decoding [10 mins]
5. MT Evaluation [10 mins]
IBM Models 1,2,3,4,5

- Models for $P(f|e)$ and $P(a|f,e)$ via $P(f,a|e)$
- There is a set of English words and the extra English word NULL
- Each English word generates and places 0 or more French words
- Any remaining French words are deemed to have been produced by NULL ("spurious words")
- Some English words may not be used at all ("zero fertility words")
IBM Model 1 parameters

\[ P(f, a|e) = P(m|\ell) \prod_i P(a_i) t(f_i|e_{a_i}) \]

\[ = \epsilon \prod_i P(a_i) t(f_i|e_{a_i}) \]

\[ = \epsilon \prod_i \frac{1}{\ell + 1} t(f_i|e_{a_i}) \]

\[ = \frac{\epsilon}{(\ell + 1)^m} \prod_i t(f_i|e_{a_i}) \]
Model 1: Word alignment learning with Expectation-Maximization (EM)

- Start with \( t(f^p|e^q) \) uniform, including \( P(f^p|\text{NULL}) \)
- For each sentence pair \((e, f)\)
  - For each French position \(i\)
    - Calculate posterior over English positions \(P(a_i | e, f)\)
      \[
P(a_i = j | f, e) = \frac{t(f_i | e_j)}{\sum_{j'} t(f_i | e_{j'})}
\]
    - Increment count of word \(f_i\) translating each word \(e_{a_i}\)
      \[- C(f_i|e_j) += P(a_i = j | f, e)\]
- Renormalize counts to give probs
  \[
t(f^p|e^q) = \frac{C(f^p|e^q)}{\sum_{f^x} C(f^x|e^q)}
\]
- Iterate until convergence
IBM Models 1,2,3,4,5

- In Model 2, the placement of a word in the French depends on where it was in the English.

- Unlike Model 1, Model 2 captures the intuition that translations should usually “lie along the diagonal.”

- A main focus of PA #1

- See Collins (2011).
Applying Model 1*

*\(P(f, a \mid e)\) can be used as a translation model or an alignment model

As translation model

\[
P(f \mid e) = \sum_a P(f, a \mid e)
\]

As alignment model

\[
P(a \mid e, f) = \frac{P(f, a \mid e)}{P(f \mid e)} = \frac{P(f, a \mid e)}{\sum_{a'} P(f, a' \mid e)}
\]

* Actually, any \(P(f, a \mid e)\), e.g., any IBM model
Summing out alignments

NULL the house
   ↘
   la maison

NULL the house
   ↘
   la maison

NULL the house
   ↘
   la maison

NULL the house
   ↘
   la maison

NULL the house
   ↘
   la maison

NULL the house
   ↘
   la maison

∑
IBM Models 1,2,3,4,5

- In Model 3, we model how many French words an English word can produce, using a concept called *fertility*

```
The proposal will not now be implemented
```

```
Les propositions ne seront pas mises en application maintenant
```

**Figure 32.3**
Alignment example.
Model 3 generative story

Mary did not slap the green witch

Mary not slap slap slap the green witch

Mary not slap slap slap NULL the green witch

Maria no dió una botefada a la verde bruja

Maria no dió una botefada a la bruja verde

Probabilities can be learned from raw bilingual text.
IBM Model 3 (from Knight 1999)

- For each word $e_j$ in English sentence, choose a fertility $\phi_j$. The choice of $\phi_j$ depends only on $e_j$, not other words or $\phi$’s: $n(\phi_j \mid e_j)$
- For each word $e_j$, generate $\phi_j$ French words. Choice of French word depends only on English word $e_j$, not on English context or any other French words.
- Permute all the French words. Each French word gets assigned absolute target position slot (1, 2, 3, etc.). Choice of French word position dependent only on absolute position of English word generating it and sentence lengths
Model 3: $P(f|e)$ parameters

- What are the parameters for this model?
- **Word translation**: $t(\text{casa} | \text{house})$
- **Spurious words**: $t(f_i | \text{NULL})$
- **Fertilities**: $n(1|\text{house})$: prob that “house” will produce 1 Spanish word whenever it appears.
- **Distortions**: $q(5|2,4,6)$: prob that word in position 2 of French translation was generated by word in position 5 of English sentence, given that 4 is length of English sentence, 6 is French length
Spurious words

• We could have \( n(3|\text{NULL}) \) (probability of there being exactly 3 spurious words in a French translation)
  – But seems wrong…
• Instead, of \( n(0|\text{NULL}) \), \( n(1|\text{NULL}) \) … \( n(25|\text{NULL}) \), have a single parameter \( p_1 \)
• After assign fertilities to non-NULL English words we want to generate (say) \( z \) French words.
• As we generate each of \( z \) words, we optionally toss in spurious French word with probability \( p_1 \)
• Probability of not adding spurious word: \( p_0 = 1 - p_1 \)
Distortion probabilities for spurious words

- Shouldn’t just have \( q(0|5,4,6) \), i.e., chance that source position for word 5 is position 0 (NULL).
- Why? These are spurious words! Could occur anywhere!! Too hard to predict
- Instead,
  - Use normal-word distortion parameters to choose positions for normally-generated French words
  - Put NULL-generated words into empty slots left over
  - If three NULL-generated words, and three empty slots, then there are 3!, or six, ways for slotting them all in
  - We’ll assign a probability of 1/6 for each way!
Model 3 parameters

- n, t, p, q
- Again, if we had complete data of English strings and step-by-step rewritings into Spanish, we could:
  - Compute \( n(0 | \text{did}) \) by locating every instance of “did”, and seeing how many words it translates to
  - \( t(\text{maison} | \text{house}) \) how many of all French words generated by “house” were “maison”
  - \( q(5 | 2, 4, 6) \) out of all times some second word is in a translation, how many times did it come from the fifth word (in sentences of length 4 and 6 respectively)?
Since we don’t have word-aligned data…

- We bootstrap alignments from incomplete data
- From a sentence-aligned bilingual corpus
  1) Assume some startup values for n, q, t, p.
  2) Use values for n, q, t, p in model 3 to work out chances of different possible alignments. Use these alignments to update values of n, q, t, p.
  3) Go to 2
- This is a more complicated case of the EM algorithm

Difficulty: Alignments are no longer independent of each other. Have to use approximate inference
Examples: translation & fertility

<table>
<thead>
<tr>
<th>the</th>
<th></th>
<th></th>
<th>not</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$t(f \mid e)$</td>
<td>$\phi$</td>
<td>$n(\phi \mid e)$</td>
<td>$t(f \mid e)$</td>
</tr>
<tr>
<td>le</td>
<td>0.497</td>
<td>1</td>
<td>0.746</td>
<td>ne</td>
</tr>
<tr>
<td>la</td>
<td>0.207</td>
<td>0</td>
<td>0.254</td>
<td>pas</td>
</tr>
<tr>
<td>les</td>
<td>0.155</td>
<td></td>
<td></td>
<td>non</td>
</tr>
<tr>
<td>l'</td>
<td>0.086</td>
<td></td>
<td></td>
<td>rien</td>
</tr>
<tr>
<td>ce</td>
<td>0.018</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cette</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

farmers

<table>
<thead>
<tr>
<th></th>
<th>$t(f \mid e)$</th>
<th>$\phi$</th>
<th>$n(\phi \mid e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>agriculteurs</td>
<td>0.442</td>
<td>2</td>
<td>0.731</td>
</tr>
<tr>
<td>les</td>
<td>0.418</td>
<td>1</td>
<td>0.228</td>
</tr>
<tr>
<td>cultivateurs</td>
<td>0.046</td>
<td>0</td>
<td>0.039</td>
</tr>
<tr>
<td>producteurs</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: idioms

<table>
<thead>
<tr>
<th>f</th>
<th>( t(f \mid e) )</th>
<th>( \phi )</th>
<th>( n(\phi \mid e) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>signe</td>
<td>0.164</td>
<td>4</td>
<td>0.342</td>
</tr>
<tr>
<td>la</td>
<td>0.123</td>
<td>3</td>
<td>0.293</td>
</tr>
<tr>
<td>tête</td>
<td>0.097</td>
<td>2</td>
<td>0.167</td>
</tr>
<tr>
<td>oui</td>
<td>0.086</td>
<td>1</td>
<td>0.163</td>
</tr>
<tr>
<td>fait</td>
<td>0.073</td>
<td>0</td>
<td>0.023</td>
</tr>
<tr>
<td>que</td>
<td>0.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hoche</td>
<td>0.054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hocher</td>
<td>0.048</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faire</td>
<td>0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>me</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>approuve</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>qui</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>un</td>
<td>0.012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faites</td>
<td>0.011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: morphology

<table>
<thead>
<tr>
<th></th>
<th>(t(f \mid e))</th>
<th>(\phi)</th>
<th>(n(\phi \mid e))</th>
</tr>
</thead>
<tbody>
<tr>
<td>devrait</td>
<td>0.330</td>
<td>1</td>
<td>0.649</td>
</tr>
<tr>
<td>devraient</td>
<td>0.123</td>
<td>0</td>
<td>0.336</td>
</tr>
<tr>
<td>devrions</td>
<td>0.109</td>
<td>2</td>
<td>0.014</td>
</tr>
<tr>
<td>faudrait</td>
<td>0.073</td>
<td></td>
<td></td>
</tr>
<tr>
<td>faut</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doit</td>
<td>0.058</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aurait</td>
<td>0.041</td>
<td></td>
<td></td>
</tr>
<tr>
<td>doivent</td>
<td>0.024</td>
<td></td>
<td></td>
</tr>
<tr>
<td>devons</td>
<td>0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>devrais</td>
<td>0.013</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IBM Models 1,2,3,4,5

• In model 4 the placement of later French words produced by an English word depends on what happened to earlier French words generated by that same English word
On Tuesday Nov. 4, earthquakes rocked Japan once again

Des tremblements de terre ont à nouveau touché le Japon mardi 4 novembre
IBM Models 1,2,3,4,5

• In model 5 they patch model 4. They make it do non-deficient alignment. That is, you can’t put probability mass on impossible things.
Alignments: linguistics

the green house

la maison verte

• There isn’t enough linguistics to explain this pattern within the translation model
• Have to depend on the language model to get it right
• That may be unrealistic
• And may be harming our translation model … and final system
IBM StatMT Translation Models

- IBM1 – lexical probabilities only
- IBM2 – lexicon plus absolute position
- HMM – lexicon plus relative position
- IBM3 – plus fertilities
- IBM4 – inverted relative position alignment
- IBM5 – non-deficient version of model 4

- All these models handle 0:1, 1:0, 1:1, 1:n alignments only

[Brown et al. 93, Vogel et al. 96]
Why all the models?

• We don’t start with aligned text, so we have to get initial alignments from somewhere.
• The alignment space has many local maxima
• Model 1 is words only, a simple model that is relatively easy and fast to train.
• The output of M1 can be a good place to start M2
  – “Starting small”. Also, it’s convex!
• The sequence of models allows a better model to be found, faster
  – The intuition is like deterministic annealing … or the pre-training done in Deep Learning
Lecture Plan

1. The IBM (Alignment) Models [30 mins]
2. Middle 10 mins: Administration, questions, catch up [10 mins]
3. Getting parallel sentences to train on [10 mins]
4. Searching for the best translation: Decoding [10 mins]
5. MT Evaluation [10 mins]
Lecture Plan

1. The IBM (Alignment) Models [30 mins]
2. Middle 10 mins: Administration, questions, catch up [10 mins]
3. Getting parallel sentences to train on [10 mins]
4. Searching for the best translation: Decoding [10 mins]
5. MT Evaluation [10 mins]
Getting Parallel Sentence Data

- **Hard way:**
  - Create your own data
  - Find and collect parallel data from web

- **Easy way:** Use existing curated data
  - Linguistic Data Consortium (LDC)
    - [http://www.ldc.upenn.edu/](http://www.ldc.upenn.edu/)
  - EuroParl/WMT:
    - [http://www.statmt.org/europarl/](http://www.statmt.org/europarl/)
    - Around 50 million words per language for “old” EU countries
Ready-to-Use Online Bilingual Data

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

+ 1m-50m words for many language pairs, e.g., all EU languages)
Tokenization (or Segmentation)

• English
  – Input (some character stream):
    "There," said Bob.
  – Output (7 “tokens” or “words”):
    "There, " said Bob.

• Chinese
  – Input (char stream):
    美国关岛国际机场及其办公室均接获
    一名自称沙地阿拉伯富商拉登等发出
    的电子邮件。
  – Output:
    美国 关岛国际 机 场 及其 办 公
    室均接获 一名 自称 沙地 阿拉 伯
    富 商拉登 等 发 出 的 电 子 电 邮。
The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchas veces. Su mujer habla con él. Los tiburones esperan.
1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchas veces.
2. Su mujer habla con él.
3. Los tiburones esperan.
1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchas veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Done by similar Dynamic Programming or EM: see FSNLP ch. 13 for details
Lecture Plan

1. The IBM (Alignment) Models [30 mins]
2. The Middle 10: Course administration, random questions, catch up, or get a head start on the back 30 [10 mins]
3. Getting parallel sentences to train on [10 mins]
4. **Searching for the best translation: Decoding** [10 mins]
5. MT Evaluation [10 mins]
Search for Best Translation

voulez – vous vous taire !
voulez – vous vous taire !

you – you you quiet !
voulez – vous vous taire !

quiet you – you you !
voulez – vous vous taire !

you shut up !
Searching for a translation

Of all conceivable English word strings, we want the one maximizing $P(e) \times P(f \mid e)$

**Exact search**

- Even if we have the right words for a translation, there are $n!$ permutations.
- We want the translation that gets the highest score under our model.
- Finding the argmax with a n-gram language model is NP-complete [Germann et al. 2001].
- Equivalent to Traveling Salesman Problem
Searching for a translation

• Several search strategies are available
  – Usually a beam search where we keep multiple stacks for candidates covering the same number of source words
  – Or, we could try “greedy decoding”, where we start by giving each word its most likely translation and then attempt a “repair” strategy of improving the translation by applying search operators (Germann et al. 2001)

• Each potential English output is called a *hypothesis*. 
Dynamic Programming Beam Search

Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

Dynamic Programming Beam Search

Each partial translation hypothesis contains:
- Last English word chosen + source words covered by it
- Next-to-last English word chosen
- Entire coverage vector (so far) of source sentence
- Language model and translation model scores (so far)

[Jelinek, 1969; Brown et al, 1996 US Patent; (Och, Ueffing, and Ney, 2001)]
Lecture Plan

1. The IBM (Alignment) Models [30 mins]
2. Middle 10 mins: Administration, questions, catch up [10 mins]
3. Getting parallel sentences to train on [10 mins]
4. Searching for the best translation: Decoding [10 mins]
5. MT Evaluation [10 mins]
Evaluating Alignments: Alignment Error Rate (Och & Ney 2000)

☐ = Sure
☐ = Possible
■ = Alignments (predicted)

\[ AER(A, S, P) = \left( 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \right) \]

\[ = \left( 1 - \frac{3 + 3}{3 + 4} \right) = \frac{1}{7} \]

Most work has used AER and we do, but it is problematic, and it’s better to use an alignment F measure (Fraser and Marcu 2007)
## Comparative results (AER)

[Och & Ney 2003]

<table>
<thead>
<tr>
<th>Model</th>
<th>Training scheme</th>
<th>0.5K</th>
<th>8K</th>
<th>128K</th>
<th>1.47M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dice</td>
<td></td>
<td>50.9</td>
<td>43.4</td>
<td>39.6</td>
<td>38.9</td>
</tr>
<tr>
<td>Dice+C</td>
<td></td>
<td>46.3</td>
<td>37.6</td>
<td>35.0</td>
<td>34.0</td>
</tr>
<tr>
<td>Model 1</td>
<td>$1^5$</td>
<td>40.6</td>
<td>33.6</td>
<td>28.6</td>
<td>25.9</td>
</tr>
<tr>
<td>Model 2</td>
<td>$1^52^5$</td>
<td>46.7</td>
<td>29.3</td>
<td>22.0</td>
<td>19.5</td>
</tr>
<tr>
<td>HMM</td>
<td>$1^5H^5$</td>
<td>26.3</td>
<td>23.3</td>
<td>15.0</td>
<td>10.8</td>
</tr>
<tr>
<td>Model 3</td>
<td>$1^52^53^3$</td>
<td>43.6</td>
<td>27.5</td>
<td>20.5</td>
<td>18.0</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^3$</td>
<td>27.5</td>
<td>22.5</td>
<td>16.6</td>
<td>13.2</td>
</tr>
<tr>
<td>Model 4</td>
<td>$1^52^53^4^3$</td>
<td>41.7</td>
<td>25.1</td>
<td>17.3</td>
<td>14.1</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^4^3$</td>
<td>26.1</td>
<td>20.2</td>
<td>13.1</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>$1^5H^54^3$</td>
<td>26.3</td>
<td>21.8</td>
<td>13.3</td>
<td>9.3</td>
</tr>
<tr>
<td>Model 5</td>
<td>$1^5H^54^35^3$</td>
<td>26.5</td>
<td>21.5</td>
<td>13.7</td>
<td>9.6</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^4^35^3$</td>
<td>26.5</td>
<td>20.4</td>
<td>13.4</td>
<td>9.4</td>
</tr>
<tr>
<td>Model 6</td>
<td>$1^5H^54^36^3$</td>
<td>26.0</td>
<td>21.6</td>
<td>12.8</td>
<td>8.8</td>
</tr>
<tr>
<td></td>
<td>$1^5H^53^4^36^3$</td>
<td>25.9</td>
<td>20.3</td>
<td>12.5</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Common software: GIZA++/Berkeley Aligner
<table>
<thead>
<tr>
<th>Foreign Original</th>
<th>Reference Translation</th>
<th>IBM4+N-grams+Stack</th>
</tr>
</thead>
<tbody>
<tr>
<td>la politique de la haine .</td>
<td>politics of hate .</td>
<td>the policy of the hatred .</td>
</tr>
<tr>
<td>nous avons signé le protocole .</td>
<td>we did sign the memorandum of agreement .</td>
<td>we have signed the protocol .</td>
</tr>
<tr>
<td>où était le plan solide ?</td>
<td>but where was the solid plan ?</td>
<td>where was the economic base ?</td>
</tr>
</tbody>
</table>

对外经济贸易合作部今天提供的数据表明，今年至十一月中国实际利用外资四百六十九点五九亿美元，其中包括外商直接投资四百点零七亿美元。

the Ministry of Foreign Trade and Economic Cooperation, including foreign direct investment 40.007 billion US dollars today provide data include that year to November china actually using foreign 46.959 billion US dollars and
MT Evaluation

• Manual (the best!?):
  – SSER (subjective sentence error rate)
  – Correct/Incorrect
  – Adequacy and Fluency (5 or 7 point scales)
  – Error categorization
  – Comparative ranking of translations

• Testing in an application that uses MT as one sub-component
  – E.g., question answering from foreign language documents
    • May not test many aspects of the translation (e.g., cross-lingual IR)

• Automatic metric:
  – WER (word error rate) – why problematic?
  – BLEU (Bilingual Evaluation Understudy)
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

**Reference (human) translation:**

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

**Machine translation:**

The American airport and its the office all receives one calls self the sand Arab rich business and so on electronic mail, which sends out; The threat will be able after public place and so on the airport to start the biochemistry attack, highly alerts after the maintenance.

**BLEU Evaluation Metric (Papineni et al, ACL-2002)**

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to match same portion of reference translation twice at a certain n-gram level (two MT words airport are only correct if two reference words airport; can’t cheat by typing out “the the the the the”)
  - Do count unigrams also in a bigram for unigram precision, etc.

- Brevity Penalty
  - Can’t just type out single word “the” (precision 1.0!)

- It was thought quite hard to “game” the system (i.e., to find a way to change machine output so that BLEU goes up, but quality doesn’t)
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

BLEU is a weighted geometric mean, with a brevity penalty factor added.

- Note that it’s precision-oriented

**BLEU4 formula**

$\exp (1.0 \times \log p1 + 0.5 \times \log p2 + 0.25 \times \log p3 + 0.125 \times \log p4 - \max(\text{words-in-reference} / \text{words-in-machine} - 1, 0))$

$p1 = 1$-gram precision

$P2 = 2$-gram precision

$P3 = 3$-gram precision

$P4 = 4$-gram precision

Note: only works at corpus level (zeroes kill it); there’s a smoothed variant for sentence-level
枪手被警方击毙。 (Foreign Original)

the gunman was shot to death by the police. (Reference Translation)

the gunman was police kill. #1
wounded police jaya of #2
the gunman was shot dead by the police. #3
the gunman arrested by police kill. #4
the gunmen were killed. #5
the gunman was shot to death by the police. #6
gunmen were killed by police ?SUB>0 ?SUB>0 #7
al by the police. #8
the ringer is killed by the police. #9
police killed the gunman. #10

green = 4-gram match (good!)
red = word not matched (bad!)
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

Reference translation 1:

Reference translation 2:

Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places.

Machine translation:

The American international airport and its office all receives one calls self the sand Arab rich business and so on electronic mail, which sends out: The threat will be able after public place and so on the airport to start the biochemistry attack, [?] highly alerts after the maintenance.

Reference translation 3:

The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden, which threatens to launch a biochemical attack on such public places as airport. Guam authority has been on alert.

Reference translation 4:

US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia. They said there would be biochemistry air raid to Guam Airport and other public places. Guam needs to be in high precaution about this matter.
Initial results showed that BLEU predicts human judgments well.

slide from G. Doddington (NIST)
Automatic evaluation of MT

• People started optimizing their systems to maximize BLEU score
  – BLEU scores improved rapidly
  – The correlation between BLEU and human judgments of quality went way, way down
  – StatMT BLEU scores now approach those of human translations but their true quality remains far below human translations
• Coming up with automatic MT evaluations has become its own research field
  – There are many proposals: TER, METEOR, MaxSim, SEPIA, our own RTE-MT
  – TERpA is a representative good one that handles some word choice variation.
• MT research requires some automatic metric to allow a rapid development and evaluation cycle.
Pots of data

• You build a model on a training set.
• Commonly, you then set further hyperparameters on another set of data, the tuning set
  • But it’s the training set for the hyperparameters
• You measure progress as you go on a dev set (development test set)
  • If you do that a lot you overfit to the dev set so it’s good to have a second dev set, dev2 set
• You evaluate and present final numbers on a test set
Pots of data

• For different reasons, the train, tune, dev, and test sets need to be completely distinct

• It is invalid to test on material you have trained on.

• If you keep running on the same evaluation set, you also begin to overfit to the evaluation set
  • Effectively you are “training” on the evaluation set ... you are learning things that do and don’t work on that particular training set.

• To get a valid measure of system performance you need another independent test set
  • Ideally, you only test on it once ... definitely very few times