Lecture 5: Machine Translation (phrases, decoding, evaluation)

Intro to NLP, CS585, Fall 2014
http://people.cs.umass.edu/~brenocon/inlp2014/
Brendan O’Connor (http://brenocon.com)

Material borrowed from Adam Lopez, Chris Manning, some combination of {Dyer, Callison-Burch, Lopez, Post}, and maybe others
• Review EM for Model 1
• Machine translation: phrase-based methods, decoding, evaluation
Word alignment models

Alignment model: ordering
- Model 1: uniform and independent
- Position movement (Model 2)
- Constraining neighboring alignments (HMM)
- One-to-many/zero/n tendencies (fertility)

\[
p(f, a \mid e) = p(f \mid e, a) \cdot p(a \mid e)
\]

Lexical translations (All IBM Models)

(Collins f,e notation)
Fancier word alignment models

\[ p(f, a \mid e) = p(f \mid e, a) \ p(a \mid e) \]

- **IBM Model 4**
  - Models fertility: \( p(\text{num e translations} \mid f \text{ word}) \)
Phrase-based MT

\[ p(f, a \mid e) = p(f \mid e, a) \ p(a \mid e) \]

\[ \downarrow \]

Phrase-to-phrase translations

- Phrases can memorize local reorderings
- State-of-the-art (currently or very recently) in industry, e.g. Google Translate
Phrase extraction for training:
Preprocess with IBM Models to predict alignments

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>open</th>
<th>the</th>
<th>box</th>
</tr>
</thead>
<tbody>
<tr>
<td>watashi</td>
<td></td>
<td>🟢</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wa</td>
<td></td>
<td>🟢</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hako</td>
<td></td>
<td></td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>wo</td>
<td></td>
<td></td>
<td>🟢</td>
<td>🟢</td>
</tr>
<tr>
<td>akemasu</td>
<td></td>
<td></td>
<td></td>
<td>🟢</td>
</tr>
</tbody>
</table>

hako wo akemasu / open the box
Chapter 25. Machine Translation

amount of transfer knowledge needed as we move up the triangl

e, from huge amounts

of transfer at the direct level (almost all knowledge is tran

rules only for parse trees

no specific transfer knowledge).

Figure 25.3 The Vauquois triangle.

In the next sections we’ll see how these algorithms address s

ome of the four trans-

ation examples shown in Fig. 25.4

English

Mary didn’t slap the green witch

⇒

Spanish

Maria Mary no dió una bofetada a la bruja verde

English

The green witch is at home this week

⇒

German

Diese Woche ist die grüne Hexe zu Hause.

English

He adores listening to music

⇒

Japanese

kare ha ongaku kiku no ga daisuki

Chinese

cheng long dao xiang gang qu

⇒

English

Jackie Chan went to Hong Kong
“Decoding”: searching for the best translation

voulez – vous vous taire !
“Decoding”: searching for the best translation

voulez – vous vous taire !

you – you you quiet !
“Decoding”: searching for the best translation

voulez – vous vous taire !

quiet you – you you !
“Decoding”: searching for the best translation

voulez – vous vous taire !

you shut up !
“Decoding”: searching for the best translation

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

**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
“Decoding”: searching for the best 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.
Maria no dio una bofetada a la bruja verde

Mary did not give a slap to the witch green
did not a slap by hag bawdy
no slap to the green witch
did not give the the witch
María no dio una bofetada a la bruja verde.

Mary did not give a slap to the witch green.

Did not a slap by hag bawdy.

No slap to the green witch.

Did not give the the witch.
Maria no dio una bofetada a la bruja verde

Mary did not give a slap to the witch green

did not a slap by hag bawdy

no slap to the green witch

did not give the the witch
Table 1: The seven-member crew includes astronauts from France and Russia.

Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
Table 1: The seven-member crew includes astronauts from France and Russia.

Scoring: Try to use phrase pairs that have been frequently observed. Try to output a sentence with frequent English word sequences.
MT Evaluation
### Illustrative translation results

<table>
<thead>
<tr>
<th>Original (Foreign)</th>
<th>Translation</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>la politique de la haine.</em></td>
<td>politics of hate</td>
<td>(Foreign Original)</td>
</tr>
<tr>
<td><em>nous avons signé le protocole.</em></td>
<td>we did sign the memorandum of agreement</td>
<td>(Foreign Original)</td>
</tr>
<tr>
<td><em>où était le plan solide ?</em></td>
<td>but where was the solid plan ?</td>
<td>(Foreign Original)</td>
</tr>
<tr>
<td><em>the policy of the hatred.</em></td>
<td>the policy of the hatred</td>
<td>(Reference Translation)</td>
</tr>
<tr>
<td><em>we have signed the protocol.</em></td>
<td>where was the economic base ?</td>
<td>(Reference Translation)</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 [?] international 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 a 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)
BLEU Evaluation Metric
(Papineni et al, ACL-2002)

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it’s precision-oriented
- BLEU4 formula
  (counts n-grams up to length 4)

\[
\exp (1.0 \times \log p_1 + \\
0.5 \times \log p_2 + \\
0.25 \times \log p_3 + \\
0.125 \times \log p_4 - \\
\max(\text{words-in-reference} / \text{words-in-machine} - 1, 0)
\]

\[p_1 = \text{1-gram precision} \]
\[P_2 = \text{2-gram precision} \]
\[P_3 = \text{3-gram precision} \]
\[P_4 = \text{4-gram precision} \]

Note: only works at corpus level (zeroes kill it);
there’s a smoothed variant for sentence-level
BLEU in Action

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 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 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 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 a [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.

\[ R^2 = 88.0\% \]

\[ R^2 = 90.2\% \]

---

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.