MT Evaluation

- **Manual (the best?):**
  - SSER (subjective sentence error rate)
  - Correct/incorrect
  - Error categorization

- **Testing in an application that uses MT as one sub-component**
  - Question answering from foreign language documents

- **Automatic metric:**
  - WER (word error rate) – why problematic?
  - BLEU (Bilingual Evaluation Understudy)

Illustrative translation results

- la politique de la haine. (Foreign Original)
- the policy of the hatred. (Reference Translation)

- nous avons signé le protocole. (Foreign Original)
- we have signed the protocol. (Reference Translation)

- où était le plan solide ? (Foreign Original)
- where was the solid plan ? (Reference Translation)

- where was the economic base ? (Reference Translation)

- 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 data actually using foreign 46.590 billion US dollars and

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

- 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 Somali Liberation Army, and
  - threatening a
  - biological/or nuclear attack, against
  - public places such as the airport.

- 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 use same portion of reference
  - translation twice (can’t cheat by typing out
  - ‘the the the’)

- Brevity penalty
  - Can’t just take out single word “the”
  - (precision 1.0)

- *** Quite hard to ‘game’ the system i.e., find a way to
  - change machine output so that BLEU goes
  - up, but quality doesn’t

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

- BLEU4 formula
  - (counts n-grams up to length 4)
  - \[
  \exp \left( \frac{1}{|S|} \sum_{i=1}^{|S|} \log p_i \right) +
  \frac{1}{|S|} \sum_{i=1}^{|S|} \log p_i +
  \frac{1}{|S|} \sum_{i=1}^{|S|} \log p_3 +
  \frac{1}{|S|} \sum_{i=1}^{|S|} \log p_4 -
  \frac{1}{|S|} \sum_{i=1}^{|S|} \log p_N -
  \frac{\sum_{i=1}^{|S|} \text{shared} \text{words in reference / words in machine} - 1}{|S|}
  \]

- Machine 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 Somali Liberation Army, and
  - threatening a
  - biological/or nuclear attack, against
  - public places such as the airport.

- BLEU in Action

- **Chinese Original:**
  - the gunman was shot to death by the police.

- **Reference:**
  - the gunman was police kill.

- **Machine translation:**
  - the gunman was police kill.

- red = word not matched (bad)
- green = 4-gram match (good)
Multiple Reference Translations

A complete translation system

Decoding for IBM Models

- Of all conceivable English word strings, find the one maximizing \( P(e) \times P(f | e) \)
- Decoding is an NP-complete challenge
  - (Knight, 1999)
- Several search strategies are available
- Each potential English output is called a hypothesis.

Dynamic Programming Beam Search
The “Fundamental Equation of Machine Translation” (Brown et al. 1993)

\[
\hat{e} = \arg\max_e P(e \mid f) \\
= \arg\max_e P(e) \times P(f \mid e) / P(f) \\
= \arg\max_e P(e) \times P(f \mid e)
\]

\[
\text{argmax } P(e \mid f) = e \\
\text{argmax } P(e) \times P(f \mid e) / P(f) = e \\
\text{argmax } P(e)^{2.4} \times P(f \mid e) \quad \ldots \text{works better!}
\]

\[
\text{Which model are you now paying more attention to?}
\]

\[
\text{argmax } P(e \mid f) = e \\
\text{argmax } P(e) \times P(f \mid e) / P(f) = e \\
\text{argmax } P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} \times \text{KS}^{3.7} \ldots
\]

\[
\text{Lots of knowledge sources vote on any given hypothesis.} \\
\text{“Knowledge source” = “feature function” = “score component”.} \\
\text{Feature function simply scores a hypothesis with a real value.} \\
\text{(May be binary, as in “e has a verb”).} \\
\text{Problem: How to set the exponent weights?}
\]

\textbf{Flaws of Word-Based MT}

- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - “real estate”, “note that”, “interested in”
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

\textbf{Phrase-Based Statistical MT}
Phrase-Based Statistical MT

- Foreign input segmented into phrases
  - "phrase" is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered

See [Koehn et al., 2003] for an intro.
This is state-of-the-art!

Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Local context is very useful for disambiguating
  - "interest rate" → 
  - "interest in" → 
- The more data, the longer the learned phrases
  - Sometimes whole sentences

How to Learn the Phrase Translation Table?

- One method: “alignment templates” (Och et al., 1999)
- Start with word alignment, build phrases from that.

IBM Models are 1-to-Many

- Run IBM-style aligner both directions, then merge:

How to Learn the Phrase Translation Table?

- One method: “alignment templates” (Och et al., 1999)
- Start with word alignment, build phrases from that.

This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.
This is the best (or “Viterbi”) alignment.

How to Learn the Phrase Translation Table?

- Collect all phrase pairs *that are consistent with the word alignment*

This word-to-word alignment is a by-product of training a translation model like IBM-Model-3.
This is the best (or “Viterbi”) alignment.

One example phrase pair
Consistent with Word Alignment

- Phrase alignment must contain all alignment points for all the words in both phrases!
- These phrase alignments are sometimes called beads

Syntax and Semantics in Statistical MT

MT Pyramid

Why Syntax?

- Need much more grammatical output
- Need accurate control over re-ordering
- Need accurate insertion of function words
- Word translations need to depend on grammatically-related words

Yamada and Knight (2001): The need for phrasal syntax

- He adores listening to music.

Syntax-based Model

- E→J Translation (Channel) Model
  
  Parse Tree (English) \( \xrightarrow{\text{Translation model}} \) Sentence (Japanese)

- Preprocess English by a parser
- Probabilistic Operations on a parse-tree
  1. Reorder child nodes
  2. Insert extra nodes
  3. Translate leaf words

\[
\begin{array}{c}
\text{Mary} \\
\text{did} \\
\text{not} \\
\text{slap} \\
\text{consistent} \\
\text{Mary} \\
\text{did} \\
\text{not} \\
\text{slap} \\
\text{inconsistent} \\
\end{array}
\]
Parse Tree(E) → Sentence (J)

1. Reorder

Parameter Table: Reorder

2. Insert

Parameter Table: Insert

3. Translate

Conditioning Feature: Child label Sequence

Conditioning Feature: Parent Label & Node Label (position) none (word selection)

Conditioning Feature= word (E) identity
Parameter Table: Translate

<table>
<thead>
<tr>
<th></th>
<th>En</th>
<th>N</th>
<th>Ja</th>
<th>N</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel</td>
<td>0.047</td>
<td>0.058</td>
<td>0.030</td>
<td>0.078</td>
<td>0.130</td>
</tr>
<tr>
<td>Plain</td>
<td>0.034</td>
<td>0.033</td>
<td>0.004</td>
<td>0.063</td>
<td>0.129</td>
</tr>
</tbody>
</table>

Note: Translation to NULL = deletion

Experiment

- Training Corpus: J-E 2K sentence pairs
- J: Tokenized by Chasen [Matsumoto, et al., 1999]
- E: Parsed by Collins Parser [Collins, 1996]
- Trained 40K Treebank, Accuracy ~90%
- E: Flatten parse tree
- To Capture word-order difference (SVO->SOV)
- EM Training: 20 iterations
- 50 mini/s (Sparc 2000MHz 1-CPU) or
- 30 sec/iter (Pentium3 700MHz 30-CPU)

Result: Alignments

<table>
<thead>
<tr>
<th>Ave. Score</th>
<th># perf sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y/K Model</td>
<td>0.582</td>
</tr>
<tr>
<td>IBM Model 5</td>
<td>0.431</td>
</tr>
</tbody>
</table>

- Ave. by 3 humans for 50 sents
- okay(1.0), not sure(0.5), wrong(0.0)
- precision only

Result: Alignment 1

Syntax-based Model

He adores listening to music
彼は音楽を聞くのが大好きです
IBM Model 3
He adores listening to music
彼は音楽を聞くのが大好きです

Result: Alignment 2

Syntax-based model

彼は拳銃を私に向けた
He aimed a revolver at me
IBM Model 3
He aimed a revolver at me
彼は拳銃を私に向けた

Result: Alignment 3

Syntax-based Model

He has unusual ability in English
彼は英語にすばるけた才能を持っている
IBM Model 3
He has unusual ability in English
彼は英語にすばるけた才能を持っている
Decoding

- Reverse direction of translation channel
  - English Parse Tree ← foreign sentence
- Use Trigram for LM
- Decoding as parsing
  - expand English grammar with model operations (reorder, insert, translate)
  - additional info (cost, reorder)

Decoding Grammar

Base English grammar
-Expanding by Model

Decoded Tree

Decoding Grammar for C/E 3M corpus

Base English grammar
-Expanding by Model

Reducing Decoder search space

- Beam search
  - Dynamic-programming parser
  - Bottom-up within beam-width (similar to [Collins 1999])
- Prune decoding grammar
  - prune by rule likelihood
  - Use extra statistics outside of model

Machine Translation

- Usable Technologies
  - "Translation memories" to aid translator
  - Low quality screening/web translators
- Technologies
  - Traditional Systran (AltaVista Babelfish, Google) is now seen as a limited success
  - Statistical MT over huge training sets is quite successful (LanguageWeaver, Microsoft, Google's future)
- Key ideas for the future
  - Statistical phrases
  - Syntax based models
  - Better language models in other respects (e.g., grammar)
  - Usably efficient decoding models (by restricting model?)