MT Evaluation
Illustrative translation results

- *la politique de la haine*. (Foreign Original)
- politics of hate. (Reference Translation)
- the policy of the hatred. (IBM4+N-grams+Stack)

- *nous avons signé le protocole*. (Foreign Original)
- we did sign the memorandum of agreement. (Reference Translation)
- we have signed the protocol. (IBM4+N-grams+Stack)

- *où était le plan solide?*. (Foreign Original)
- but where was the solid plan? (Reference Translation)
- where was the economic base? (IBM4+N-grams+Stack)

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
  – 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)
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.

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 the the”)

Brevity penalty
- Can’t just type 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)
- Caveat: More recently, people seem to have been finding ways....
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
  (counts n-grams up to length 4)

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

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

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.

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.
BLEU Tends to Predict Human Judgments

slide from G. Doddington (NIST)
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 NP hard
  – (Knight, 1999)

• Several search strategies are available
  – Usually a beam search where we keep multiple stacks for candidates covering the same number of source words

• 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)

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)

The “Fundamental Equation of Machine Translation” (Brown et al. 1993)

\[ \hat{e} = \arg\max_e P(e | f) \]

\[ = \arg\max_e P(e) \times P(f | e) / P(f) \]

\[ = \arg\max_e P(e) \times P(f | e) \]
What StatMT people do in the privacy of their own homes

$$\arg\max \: P(e \mid f) = e$$

$$\arg\max \: P(e) \times P(f \mid e) / P(f) = e$$

$$\arg\max \: P(e)^{2.4} \times P(f \mid e) \quad \text{… works better!}$$

Which model are you now paying more attention to?
What StatMT people do in the privacy of their own homes

\[
\text{argmax } P(e \mid f) = e
\]

\[
\text{argmax } P(e) \times P(f \mid e) / P(f) = e
\]

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

Rewards longer hypotheses, since these are ‘unfairly’ punished by \( P(e) \).
What StatMT people do in the privacy of their own homes

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

Lots of knowledge sources vote on any given hypothesis.

“Knowledge source” = “feature function” = “score component”.

Feature function simply scores a hypothesis with a real value.

(May be binary, as in “e has a verb”).

**Problem:** How to set the exponent weights?
(We look at one way later: maxent models.)
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
Alignments: linguistics

the green house

\[ \times \]

la maison verte

• There isn’t enough linguistics to explain this in the translation model ... have to depend on the language model ... that may be unrealistic ... and may be harming our translation model
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
  – $P(\text{to the conference} \mid \text{zur Konferenz})$
  – $P(\text{into the meeting} \mid \text{zur Konferenz})$
• Phrases are probabilistically re-ordered
This is the 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.

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dió</th>
<th>una bofetada a</th>
<th>la bruja verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td></td>
<td>did</td>
<td></td>
<td></td>
</tr>
<tr>
<td>did</td>
<td></td>
<td>not</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td>slap</td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td></td>
<td>the</td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td>green</td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td>witch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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?

- 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.
IBM Models are 1-to-Many

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

  - $E \rightarrow F$ best alignment
  - $F \rightarrow E$ best alignment

  ![Diagram showing the process of merging alignments](image)

  **MERGE**

Union or intersection or cleverer algorithm
How to Learn the Phrase Translation Table?

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

```
<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dió</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>did</td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

- 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
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.

彼は音楽を聞くのが大好きです
Kare ha ongaku wo kiku no ga daisuki desu
Syntax-based Model

• E→J Translation (Channel) Model

Parse Tree (English) ⇒ 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
Kare ha ongaku wo kiku no ga daisuki desu
1. Reorder

P(PRP VB1 VB2→ PRP VB2 VB1 ) = 0.723
P(VB TO → TO VB ) = 0.749
P(TO NN → NN TO ) = 0.893

Conditioning Feature = Child label Sequence
## Parameter Table: Reorder

| Original Order | Reordering          | $P(\text{reorder} | \text{original})$ |
|----------------|---------------------|-------------------|
| PRP VB1 VB2    | PRP VB1 VB2         | 0.074             |
|                | PRP VB2 VB1         | 0.723             |
|                | VB1 PRP VB2         | 0.061             |
|                | VB1 VB2 PRP         | 0.037             |
|                | VB2 PRP VB1         | 0.083             |
|                | VB2 VB1 PRP         | 0.021             |
| VB TO          | VB TO               | 0.107             |
|                | TO VB               | 0.893             |
| TO NN          | TO NN               | 0.251             |
|                | NN TO               | 0.749             |
2. Insert

\[
P(\text{none}|\text{TOP-VB}) = 0.735
\]
\[
P(\text{right}|\text{VB-PRP}) \times P(\text{ha}) = 0.652 \times 0.219
\]
\[
P(\text{right}|\text{VB-VB}) \times P(\text{ga}) = 0.252 \times 0.062
\]
\[
P(\text{none}|\text{TO-TO}) = 0.900
\]

Conditioning Feature = Parent Label & Node Label (position)
none (word selection)
## Parameter Table: Insert

<table>
<thead>
<tr>
<th>Parent label node level</th>
<th>TOP</th>
<th>VB</th>
<th>VB</th>
<th>TO</th>
<th>TO</th>
<th>TO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VB</td>
<td></td>
<td></td>
<td>VB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (none)</td>
<td>0.735</td>
<td>0.687</td>
<td>0.344</td>
<td>0.700</td>
<td>0.900</td>
<td>0.800</td>
</tr>
<tr>
<td>P (left)</td>
<td>0.004</td>
<td>0.061</td>
<td>0.004</td>
<td>0.030</td>
<td>0.003</td>
<td>0.096</td>
</tr>
<tr>
<td>P (right)</td>
<td>0.260</td>
<td>0.252</td>
<td>0.652</td>
<td>0.261</td>
<td>0.097</td>
<td>0.104</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>W</th>
<th>P (insert-w)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ha</td>
<td>0.219</td>
</tr>
<tr>
<td>ta</td>
<td>0.131</td>
</tr>
<tr>
<td>wo</td>
<td>0.099</td>
</tr>
<tr>
<td>no</td>
<td>0.094</td>
</tr>
<tr>
<td>ni</td>
<td>0.080</td>
</tr>
<tr>
<td>te</td>
<td>0.078</td>
</tr>
<tr>
<td>ga</td>
<td>0.062</td>
</tr>
<tr>
<td>desu</td>
<td>0.0007</td>
</tr>
</tbody>
</table>
3. Translate

Conditioning Feature = word (E) identity

P (he → kare) = 0.952
P (music → ongaku) = 0.900
P (to → wo) = 0.038
P (listening → kiku) = 0.333
P (adore → daisuki) = 1.000
### Parameter Table: Translate

<table>
<thead>
<tr>
<th></th>
<th>adores</th>
<th>he</th>
<th>listening</th>
<th>music</th>
<th>to</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>adores</td>
<td>he</td>
<td>listening</td>
<td>music</td>
<td>to</td>
</tr>
<tr>
<td>J</td>
<td>daisuki 1.000</td>
<td>kare 0.952</td>
<td>kiku 0.333</td>
<td>ongaku 0.900</td>
<td>ni 0.216</td>
</tr>
<tr>
<td></td>
<td>NULL 0.016</td>
<td>NULL 0.016</td>
<td>kii 0.333</td>
<td>NULL 0.204</td>
<td>to 0.133</td>
</tr>
<tr>
<td></td>
<td>nani 0.005</td>
<td>mi 0.333</td>
<td>naru 0.100</td>
<td>no 0.046</td>
<td></td>
</tr>
<tr>
<td></td>
<td>da 0.003</td>
<td></td>
<td></td>
<td>wo 0.038</td>
<td></td>
</tr>
<tr>
<td></td>
<td>shi 0.003</td>
<td></td>
<td></td>
<td></td>
<td></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, 1999]
  --- Trained: 40K Treebank, Accuracy: ~90%
- E: Flatten parse tree
  --- To Capture word-order difference (SVO->SOV)
- EM Training: 20 Iterations
  --- 50 min/iter (Sparc 200Mhz 1-CPU) or
  --- 30 sec/iter (Pentium3 700Mhz 30-CPU)
Result: Alignments

<table>
<thead>
<tr>
<th></th>
<th>Ave. Score</th>
<th># perf sent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y/K Model</td>
<td>0.582</td>
<td>10</td>
</tr>
<tr>
<td>IBM Model 5</td>
<td>0.431</td>
<td>0</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 2

Syntax-based model

He aimed a revolver at me

彼は銃を私に向けた
Kare ha kenju wo watashi ni muke ta

IBM Model 3

He aimed a revolver at me

彼は銃を私に向けた
Result: Alignment 3

Syntax-based Model

He has unusual ability in English

彼 は 英語 に ずばぬけ た 才能 を 持っ て い る
Kare ha eigo ni zubanuke ta sainou wo mottu te iru

IBM Model 3

He has unusual ability in English

彼 は 英語 に ずばぬけ た 才能 を 持っ て い る
Machine Translation Summary

• **Usable Technologies**
  – “Translation memories” to aid translator
  – Low quality screening/web translators

• **Technologies**
  – Traditional: Systran (Altavista Babelfish, what you got till mid-2006 on Google) is now seen as a limited success
  – Statistical MT over huge training sets is successful (ISI/LanguageWeaver, Microsoft, Google)

• **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?)