Illustrative translation results

- le politique de la haine. (Foreign Original)
- the policy of the hatred. (Reference Translation)
- nous avons signé le protocole. (Foreign Original)
- we did sign the memorandum of agreement. (Reference Translation)
- où était le plan solide ? (Foreign Original)
- but where was the solid plan? (Reference Translation)
- where was the economic base? (IBM+N-grams+Stack)

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

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
  - Question answering from foreign language documents

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

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

- N-gram precision (score is between 0 & 1)
  - 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”)

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

It was thought quite hard to “game” the system  
- to find a way to change machine output so that BLEU goes up, but quality doesn’t

BLEU in Action

<table>
<thead>
<tr>
<th>Reference (human) translation:</th>
<th>Foreign Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>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 sand police, please such as the airport.</td>
<td>the police</td>
</tr>
<tr>
<td>An 4-gram is an sequence of 4 words.</td>
<td>police</td>
</tr>
<tr>
<td>Can’t just type out single word “the” (precision 1.0)</td>
<td>police</td>
</tr>
<tr>
<td>It was thought quite hard to “game” the system.</td>
<td>police</td>
</tr>
<tr>
<td>to find a way to change machine output so that BLEU goes up, but quality doesn’t.</td>
<td>police</td>
</tr>
</tbody>
</table>

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 sand police, please such as the airport. The threat will be able after public place and so on. The airport to start the high state of alert. The economic base.

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

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 sand police, please such as the airport. The threat will be able after public place and so on. The airport to start the high state of alert. The economic base.

Note: only works at corpus level (zeros kill it); there’s a smoothed variant for sentence-level
Multiple Reference Translations

Quiz question!
MT Hypothesis: the gunman was shot dead by police.
– Ref 1: The gunman was shot to death by the police.
– Ref 2: The cops shot the gunman dead.

• What is the:
  – Unigram precision?
  – Trigram precision?
Note: punctuation tokens are counted in calculation

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
  – METEOR is a representative good one that handles some word choice variation.
• MT research really requires some automatic metric to allow a rapid development and evaluation cycle.

MT: The early history (1950s)
• Early attempts
  – Foundational work on automata, formal semantics
• First
• MT
• Limited
• Probability

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)

Dynamic Programming Beam Search

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

$\hat{e} = \arg\max P(e | f)$

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

$= \arg\max P(e) \times P(f | e)$

Which model are you now paying more attention to?

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

$$\arg\max_e P(e)^{1.9} \times P(f \mid e) \times 1.1^{\text{length}(e)} \times KS^{3.7} \ldots$$

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 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

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

- 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 J&M or Lopez 2008 for an intro.

This is still pretty much 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.

Mary did not slap the green witch

Maria no dió una bofetada a la bruja verde

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→F best alignment

F→E best alignment

Union or intersection or cleverer algorithm

How to Learn the Phrase Translation Table?

• Collect all phrase pairs that are consistent with the 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

**Syntax-based Model**

- E→J Translation (Channel) Model
  - Parse Tree (English) \[\Rightarrow\] Translation model \[\Rightarrow\] 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

**Parse Tree(E) → Sentence (J)**

1. Reorder

**Parameter Table: Reorder**

<table>
<thead>
<tr>
<th>Original Order</th>
<th>Reordering</th>
<th>Reorder/Original</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP VB1 VB2 VB3</td>
<td>0.774</td>
<td>VB1 VB2 PRP VB3</td>
</tr>
<tr>
<td>PRP VB2 VB1</td>
<td>0.723</td>
<td>VB1 PRP VB2 VB3</td>
</tr>
<tr>
<td>VB1 PRP VB2</td>
<td>0.011</td>
<td>VB2 VB1 PRP VB3</td>
</tr>
<tr>
<td>VB2 PRP VB1</td>
<td>0.083</td>
<td>VB2 VB1 PRP VB3</td>
</tr>
<tr>
<td>VB2 VB1 PRP</td>
<td>0.021</td>
<td>VB2 VB1 PRP VB3</td>
</tr>
<tr>
<td>VB TO</td>
<td>0.856</td>
<td>VB TO VB VB2 VB1</td>
</tr>
<tr>
<td>TO VB</td>
<td>0.893</td>
<td>TO VB VB2 VB1 VB3</td>
</tr>
<tr>
<td>TO NN</td>
<td>0.251</td>
<td>TO NN VB2 VB1 VB3</td>
</tr>
<tr>
<td>NN TO</td>
<td>0.749</td>
<td>NN TO VB2 VB1 VB3</td>
</tr>
</tbody>
</table>

Conditioning Feature = Child label Sequence

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

- He adores listening to music.

\[Kare \ ha \ ongaku \ wo \ kiku \ no \ ga \ daisuki \ desu\]
2. Insert

\[
P(\text{none}(\text{TOP-VB}) = 0.735
\]
\[
P(\text{right}(\text{VB-PRP}) * P(\text{ha}) = 0.652 * 0.219
\]
\[
P(\text{right}(\text{VB-VB}) * P(\text{ga}) = 0.262 * 0.062
\]
\[
P(\text{none}(\text{TOP-TOP}) = 0.900
\]
Conditioning Feature = Parent Label & Node Label (position)
none (word selection)

---
---
---
---
---

3. Translate

\[
P(\text{ha} \rightarrow \text{kare}) = 0.952
\]
\[
P(\text{music} \rightarrow \text{ongaku}) = 0.000
\]
\[
P(\text{to} \rightarrow \text{vo}) = 0.038
\]
\[
P(\text{listening} \rightarrow \text{kiku}) = 0.333
\]
\[
P(\text{adore} \rightarrow \text{daisuki}) = 1.000
\]
Conditioning Feature = word (E) identity

---
---
---
---
---

## 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>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 2

Result: Alignment 3

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 of the present/future**
  - Statistical phrase based models
  - Syntax based models
  - Better language models (e.g., bigger, using grammar)
  - Better decoding models (e.g., by restricting model?)