Machine Translation Systems

Christopher Manning
CS224N / Ling 284
[Based on slides by Kevin Knight, Dan Klein, Dan Jurafsky]
MT Evaluation
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 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
  (counts n-grams up to length 4)

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

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

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.

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.
Initial results showed that BLEU predicts human judgments well

R^2 = 88.0%

R^2 = 90.2%

slide from G. Doddington (NIST)
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 but not sentence boundary tokens.
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 really requires some automatic metric to allow a rapid development and evaluation cycle.
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.
Search for Best Translation

voulez – vous vous taire !
Search for Best Translation

voulez – vous vous taire !

you – you you quiet !
Search for Best Translation

voulez – vous vous taire !

quiet you – you you !
Search for Best Translation

voulez-vous vous vous taire !

you shut up !
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;
(Och, Ueffing, and Ney, 2001]
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)

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

$$\text{argmax } P(e | f) = e$$

$$\text{argmax } P(e) \times P(f | e) / P(f) = e$$

$$\text{argmax } P(e)^{1.9} \times P(f | e) \quad \ldots \text{ works better!}$$

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

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

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

\[
\text{argmax } P(e)^{1.9} \times P(f \mid e) \times 1.1^{\text{length}(e)}
\]

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

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

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
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

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?

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

\[
\begin{array}{cccccc}
\text{Maria} & \text{no} & \text{dió} & \text{una} & \text{bofetada} & \text{a} & \text{la} & \text{bruja} & \text{verde} \\
\hline
\text{Mary} & & & & & & & \\
\text{did} & & & & & & & \\
\text{not} & & & & & & & \\
\text{slap} & & & & & & & \\
\text{the} & & & & & & & \\
\text{green} & & & & & & & \\
\text{witch} & & & & & & & \\
\end{array}
\]

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.

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dió</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>did</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>slap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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.
IBM Models are 1-to-Many

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

E→F best alignment

F→E best alignment

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*

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

- A certain phrase pair (f-f-f, e-e-e) may appear many times across the bilingual corpus.

- No EM training

- Just relative frequency:

\[
P(f-f-f \mid e-e-e) = \frac{\text{count}(f-f-f, e-e-e)}{\text{count}(e-e-e)}
\]
### Phrase-Based Translation

<table>
<thead>
<tr>
<th></th>
<th>7人</th>
<th>中包括</th>
<th>来自</th>
<th>法国</th>
<th>和</th>
<th>俄罗斯</th>
<th>的</th>
<th>宇航</th>
<th>员</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>the</strong></td>
<td>7 people</td>
<td>including</td>
<td>by some</td>
<td><strong>and</strong></td>
<td>the russian</td>
<td><strong>the</strong></td>
<td>the astronauts</td>
<td>,</td>
<td></td>
</tr>
<tr>
<td><strong>it</strong></td>
<td>7 people included</td>
<td>by france</td>
<td>and the</td>
<td>russian</td>
<td><strong>international</strong></td>
<td><strong>astronautical</strong></td>
<td>of rapporteur</td>
<td>.</td>
<td></td>
</tr>
<tr>
<td><strong>this</strong></td>
<td>7 out</td>
<td>including the</td>
<td>from</td>
<td>the french</td>
<td><strong>and</strong></td>
<td>the russian</td>
<td><strong>the</strong></td>
<td>fifth</td>
<td>.</td>
</tr>
<tr>
<td><strong>these</strong></td>
<td>7 among</td>
<td>including from</td>
<td>the french</td>
<td>and</td>
<td>of the russian</td>
<td><strong>of</strong></td>
<td>space</td>
<td><strong>members</strong></td>
<td>.</td>
</tr>
<tr>
<td><strong>that</strong></td>
<td>7 persons</td>
<td>including from the</td>
<td>of france</td>
<td>and to</td>
<td>russian</td>
<td><strong>of</strong></td>
<td>the aerospace</td>
<td><strong>members</strong></td>
<td>.</td>
</tr>
<tr>
<td>7 include</td>
<td>from the</td>
<td>of france and</td>
<td>russian</td>
<td><strong>astronauts</strong></td>
<td>.</td>
<td><strong>the</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 numbers include</td>
<td>from france</td>
<td>and russian</td>
<td><strong>of astronauts who</strong></td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 populations include</td>
<td>those from france</td>
<td>and russian</td>
<td><strong>astronauts</strong></td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 deportees included</td>
<td>come from</td>
<td>france <strong>and</strong></td>
<td>russian</td>
<td><strong>in</strong></td>
<td><strong>astronautical</strong></td>
<td><strong>personnel</strong></td>
<td><strong>;</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 philtrum</td>
<td>including those from</td>
<td><strong>france and</strong></td>
<td>russian</td>
<td><strong>a space</strong></td>
<td><strong>astronaut</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>include</strong></td>
<td>came from</td>
<td><strong>france and russian</strong></td>
<td>by cosmonauts</td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>include</strong></td>
<td>representatives from</td>
<td><strong>france and</strong></td>
<td>russian</td>
<td><strong>cosmonauts</strong></td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>include</strong></td>
<td>came from france</td>
<td>and russian’s</td>
<td><strong>cosmonauts</strong></td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>includes</strong></td>
<td>coming from</td>
<td><strong>france and</strong></td>
<td>russian’s</td>
<td><strong>’s astronaut</strong></td>
<td><strong>navigation</strong></td>
<td><strong>member</strong></td>
<td>.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>french</strong></td>
<td><strong>and</strong></td>
<td>russian</td>
<td><strong>astronauts</strong></td>
<td>special rapporteur</td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>and</strong></td>
<td>russian’s</td>
<td><strong>;</strong></td>
<td><strong>and</strong></td>
<td><strong>russia</strong></td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>and</strong></td>
<td>russian</td>
<td><strong>;</strong></td>
<td>rapporteur</td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>and</strong></td>
<td>russian</td>
<td><strong>;</strong></td>
<td>rapporteur</td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>and</strong></td>
<td>russian</td>
<td><strong>;</strong></td>
<td>or</td>
<td>russian’s</td>
<td><strong>;</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** #11# 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.
Phrase-Based Translation

这 7 人 中包括 来自 法国 和 俄罗斯 的 宇航 员 .

<table>
<thead>
<tr>
<th>the 7 people including by some and the russian the astronauts</th>
</tr>
</thead>
<tbody>
<tr>
<td>it 7 people included by france and the russian international astronomical of rapporteur .</td>
</tr>
<tr>
<td>these 7 among including from the french and the russian the fifth</td>
</tr>
<tr>
<td>that 7 persons including from the of french and to russian of the aerospace members</td>
</tr>
<tr>
<td>7 include from the of french and russian astronauts</td>
</tr>
<tr>
<td>7 numbers include from france and russian of astronauts who</td>
</tr>
<tr>
<td>7 populations include those from france and russia astronauts</td>
</tr>
<tr>
<td>7 deportees included come from france and russia in astronomical personnel</td>
</tr>
<tr>
<td>7 philtrum including those from france and russia a space member</td>
</tr>
<tr>
<td>including representatives from france and the russia astronaut</td>
</tr>
<tr>
<td>include came from france and russia by cosmonauts</td>
</tr>
<tr>
<td>include came from france and russia's cosmonaut</td>
</tr>
<tr>
<td>includes coming from french and russia's cosmonaut</td>
</tr>
<tr>
<td>french and russian 's astronavigation member .</td>
</tr>
<tr>
<td>french and russia's astronauts special rapporteur</td>
</tr>
<tr>
<td>, and russia rapporteur</td>
</tr>
<tr>
<td>, and russia rapporteur</td>
</tr>
<tr>
<td>, and russia</td>
</tr>
<tr>
<td>or russia's</td>
</tr>
</tbody>
</table>

Table 1: #11# 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>
<thead>
<tr>
<th>This</th>
<th>7 people</th>
<th>included</th>
<th>by some</th>
<th>and</th>
<th>the Russian</th>
<th>the</th>
<th>astronauts</th>
</tr>
</thead>
<tbody>
<tr>
<td>that</td>
<td>7 people included</td>
<td>by French</td>
<td>and the Russian</td>
<td>international</td>
<td>astronomical</td>
<td>of</td>
<td>rapporteur</td>
</tr>
<tr>
<td>these</td>
<td>7 among</td>
<td>including from</td>
<td>the French</td>
<td>and the Russian</td>
<td>the fifth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td>7 persons</td>
<td>including from</td>
<td>the French and</td>
<td>of the Russian</td>
<td>of</td>
<td>space members</td>
<td></td>
</tr>
<tr>
<td>7 include</td>
<td>from the</td>
<td>of French and</td>
<td>Russian</td>
<td>astronauts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 numbers include</td>
<td>from France</td>
<td>and Russian</td>
<td>of astronauts who</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 populations include</td>
<td>those from French</td>
<td>and Russian</td>
<td>astronauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 deportees include</td>
<td>come from</td>
<td>French and Russia</td>
<td>in</td>
<td>astronomical personnel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 philltrum</td>
<td>including those from</td>
<td>French and</td>
<td>Russia</td>
<td>a space member</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include</td>
<td>came from</td>
<td>French and</td>
<td>Russia</td>
<td>by cosmonauts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include</td>
<td>came from</td>
<td>French and</td>
<td>Russia’s</td>
<td>cosmonauts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes</td>
<td>coming from</td>
<td>French and</td>
<td>Russia’s</td>
<td>cosmonaut</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>french</td>
<td>and</td>
<td>Russian</td>
<td>’s</td>
<td>astronaut navigation member</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>french</td>
<td>and</td>
<td>Russian</td>
<td>astronauts</td>
<td>special rapporteur</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: #11# 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.
### Phrase-Based Translation

<table>
<thead>
<tr>
<th>This</th>
<th>7人</th>
<th>中包括</th>
<th>来自</th>
<th>法国</th>
<th>和</th>
<th>俄罗斯</th>
<th>的</th>
<th>宇航员</th>
<th>.</th>
</tr>
</thead>
<tbody>
<tr>
<td>the 7 people</td>
<td>including by some</td>
<td>and the Russian</td>
<td>the astronauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>it 7 people included by France</td>
<td>and the Russian</td>
<td>international astronomical of rapporteur</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>these 7 among including from the French and of the Russian</td>
<td>the fifth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that 7 persons including from French and to Russian</td>
<td>of the ensembles members</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 include from the French and Russian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 numbers include from France and Russian</td>
<td>of astronauts who</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 populations include those from French and Russian</td>
<td>astronauts .</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 deportees included come from France and Russia</td>
<td>in astronomical</td>
<td>personnel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 plhiatm including those from France and Russia</td>
<td>a space member</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include came from France and Russia</td>
<td>by cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include representatives from French and Russia</td>
<td>cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>include came from France and Russia ’s cosmonauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>includes coming from French and Russian</td>
<td>'s astronaut navigation member</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>french and Russian</td>
<td>astronauts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Scoring:
- Try to use phrase pairs that have been frequently observed.
- Try to output a sentence with frequent English word sequences.

Table 1: #11# the seven-member crew includes astronauts from France and Russia.
Syntax and Semantics in Statistical MT
MT Pyramid

interlingua

semantics

syntax

phrases

words

SOURCE

TARGET
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
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

彼は拳銃を私に向けた
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

彼は英語にずばぬけた才能を持つている
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つている
MT Applications

Christopher Manning
CS 224N
2009
MT: The early history (1950s)

- Early NLP (Machine Translation) on machines less powerful than pocket calculators
- Foundational work on automata, formal languages, probabilities, and information theory
- First speech systems (Davis et al., Bell Labs)
- MT heavily funded by military, but basically just word substitution programs
- Little understanding of natural language syntax, semantics
- Problem soon appeared intractable
MT Applications: 1. Traditional

- Traditional scenario:
  - Documents had to be translated for your company/organization. Document production for organization
  - Generally, the quality/accuracy demands are high
  - High cost
    - Though most of it is now done as outsourced piecework
- MT tends to be ineffective: The cost of post-translation error correction is too high
- Main technology in the game: translation memory/translation workbench/terminology management
  - E.g., TRADOS.
    - Very slowly, MT technology is starting to be incorporated, but most of the action is in terminology lexicon management
**T. I. E. Japan Co., Ltd - T. I. E. Japan Co., Ltd**

*Accuracy. Reliability. Reasonable price.*

Minato-ku, Tokyo, Japan / Native in: Japanese

<table>
<thead>
<tr>
<th>Profile</th>
<th>Company profile</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freelancer and outsourcer</td>
<td>Translation, Website localization, Software localization, Desktop publishing</td>
<td></td>
</tr>
<tr>
<td><strong>Expertise</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specializes in:</td>
<td>Mathematics &amp; Statistics</td>
<td></td>
</tr>
<tr>
<td>Business/Commerce (general)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mechanics / Mech Engineering</td>
<td>Journalism</td>
<td></td>
</tr>
<tr>
<td>IT (Information Technology)</td>
<td>Internet, e-Commerce</td>
<td></td>
</tr>
<tr>
<td>Electronics / Elect Eng</td>
<td>Telecom(munications)</td>
<td></td>
</tr>
<tr>
<td>Computers (general)</td>
<td>Computers: Systems, Networks</td>
<td></td>
</tr>
<tr>
<td><strong>Rates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English to Japanese - Rates: 0.10 - 0.06 USD per word / 40 - 20 USD per hour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese to Japanese - Rates: 0.13 - 0.08 USD per word / 45 - 25 USD per hour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Japanese to Portuguese - Rates: 0.13 - 0.08 USD per word / 45 - 25 USD per hour</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish to Japanese - Rates: 0.13 - 0.08 USD per word / 45 - 25 USD per hour</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>KudoZ activity</strong></td>
<td>Questions answered: 2, Questions asked: 0 Easy / 0 PRO</td>
<td></td>
</tr>
<tr>
<td><strong>Translation education</strong></td>
<td>Univ. de Sao Paulo</td>
<td></td>
</tr>
<tr>
<td><strong>ProZ.com Certified</strong></td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>PRO certificate(s)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Credentials</strong></td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Memberships</strong></td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td><strong>Software</strong></td>
<td>Adobe Acrobat, Adobe Photoshop, Microsoft Excel, Microsoft Word, Pagemaker, Powerpoint</td>
<td></td>
</tr>
<tr>
<td><strong>Website</strong></td>
<td><a href="http://www.logos-net.com">http://www.logos-net.com</a></td>
<td></td>
</tr>
<tr>
<td><strong>CV/Resume</strong></td>
<td>N/A</td>
<td>CV/Resume</td>
</tr>
</tbody>
</table>

**Languages** - Japanese / Brazilian Portuguese / Spanish / English

English-Japanese
Brazilian Portuguese-Japanese, Japanese-Brazilian Portuguese
English-Brazilian Portuguese
Spanish-Japanese

**Mother tongue**: Japanese
Trados is relatively pricey (high hundreds for PC versions, thousands for server version); seen as necessary productivity tool (Photoshop for translators)
MT Applications: 2. Web

• Web applications:
  – Dominant scenario: User-initiated translation
    • Crucial difference: The quality doesn’t have to be great. The user is usually okay if they can understand the gist of what is going on
  – Second scenario
    • Somehow on the web people will accept medium quality results. Accessible information is better than no information

• MT is saved!!! “It’s the web, stupid.”
  • (But is there money in it?)
Altavista
BabelFish
1997:
Free, automatic translation for the masses.
Revolutionary.

But, what was the underlying technology?
SYSTRAN.

MacOS Dashboard?
SYSTRAN

Google until 2006?
SYSTRAN
SYSTRAN, founded by Dr. Peter Toma in 1968[^1], is one of the oldest machine translation companies. SYSTRAN has done extensive work for the United States Department of Defense and the European Commission.

SYSTRAN provides the technology for Yahoo! and AltaVista's (Babel Fish) among others, but use of it was ended (circa 2007) for all of the language combinations offered by Google's language tools[^2].

SYSTRAN is used by the Dashboard Translation widget in Mac OS X.

Commercial versions of SYSTRAN operate with operating systems Microsoft Windows (including Windows Mobile), Linux and Solaris.

### History

With its origin in the Georgetown machine translation effort, SYSTRAN was one of the few machine translation systems to survive the major decrease of funding after the ALPAC Report of the mid-1960s. The company was established in La Jolla, California to work on translation of Russian to English text for the United States Air Force during the Cold War. Large numbers of Russian scientific and technical documents were translated using SYSTRAN under the auspices of the USAF Foreign Technology Division (later the National Air and Space Intelligence Center) at Wright-Patterson Air Force Base, Ohio. The quality of the translations, although only approximate, was usually adequate for understanding content.

The company was sold in 1986 to the Gachot family, based in Paris, France, and is now traded publicly by the French stock exchange. It has a main office at the Grande Arche in La Defense and maintains a secondary office in La Jolla, San Diego, California.

During the dot com boom, language industry started a new era, and Systran entered into agreements with a number of translation integrators, the most successful of these being WorldLingo, today many millions of people use Systran's Translation Engines through WorldLingo.
SYSTRAN Translation Software

SYSTRAN's wide range of translation software products and solutions will help you create multilingual documents and understand foreign language content in real-time.

- **Home & Home Office**
  - Web Translator
  - Home Translator
  - Office Translator
  - Mobile Translator
  - All products

- **Small Business**
  - Business Translator
  - Premium Translator
  - Enterprise Server 6
  - SYSTRANLinks

- **Enterprise Solutions**
  - Premium Translator
  - Business Translator
  - Enterprise Server 5
  - SYSTRANLinks
  - Professional Services

**Buy Desktop Product Online**

An international Benchmark in Translation Software

SYSTRAN is the market-leading provider of language translation software products and solutions for desktop, enterprise and Internet that facilitate communication in 52 language combinations and in 20 vertical domains. With over four decades of expertise in Machine Translation dedicated to multilingual translation, SYSTRAN's software is the choice of leading global corporations, Internet portals and public agencies including the US intelligence Community and the European Commission.

What is automatic translation?

Automatic translation is a translation produced by state-of-the-art technology, without the intervention of human translators. Automatic translation is also often referred to as "Machine Translation".
SYSTEM REQUIREMENTS

Client Server 32-Bit Platform
- Processor: 2.8 GHz Pentium 4
- RAM: 4 GB

Client Server 64-Bit Platform
- OS: Windows 2003 Server, Windows 2000 Advanced Server (SP3+) or Linux Fedora 8 or similar (contact us for details)
- Processor: 2.8 GHz Intel Xeon
- RAM: 8 GB

RUN TIME TRANSLATION PROCESS

TRANSLATION PARAMETERS
(Bilingual Data)

<table>
<thead>
<tr>
<th>Mejores</th>
<th>19%</th>
<th>7.3%</th>
<th>2.3%</th>
<th>2.2%</th>
<th>2%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Better</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outperform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mundo</th>
<th>4.7%</th>
<th>3.7%</th>
<th>3.1%</th>
<th>2.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Globe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Globalizing</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worlds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LANGUAGE MODEL
(Monolingual Data)

| The best companies in the world | 27% |
| Best companies in the globe    | 17% |
| World's best companies        | 11% |
| Better companies in the world  | 5.1% |
| The world companies are best   | 4.8% |
Language Tools

Search across languages

Type a search phrase in your own language to easily find pages in another language. We'll translate the results for you to read.

Search for: 

My language: English  Search pages written in: Spanish

Translate and Search

Tip: Use advanced search to restrict your search by language and country without translating your search phrase.

Translate text

hola

Albanian
Arabic
Bulgarian
Catalan
Chinese (Simplified)
Chinese (Traditional)
Croatian
Czech
Danish
Dutch
English
Estonian
Filipino
Finnish
French
Galician

Translate a web page

http://

Translate

Translate

Translate

Translate
CORREGIR: es que recibas un mensaje de error cuando intentas cerrar una aplicación basada en formularios de Windows cuando la aplicación se ejecuta en .NET Framework 2.0.
Corrige un problema que se puede producir cuando intenta cerrar una aplicación basada en formularios de Windows que utiliza la información sobre herramientas. Es posible que reciba un "System.NullReferenceException: objetos de referencia no se...

Más Descargas...

Solución de problemas – .NET Framework 2.0

- El control de usuario de ActiveX no está visible en el diseñador de formularios cuando se vuelve a abrir un proyecto existente.
  Explica que el control de usuario de ActiveX no es visible en el diseñador de Windows Forms cuando se vuelve a abrir un proyecto existente. Proporciona un método para solucionar este problema.

- No puede establecer el foco a los controles secundarios de un control de usuario mediante la tecla TAB.
  Explica que no establece el foco a los controles secundarios de un control de usuario mediante la tecla TAB. Proporciona una solución para solucionar este problema.

- ERROR: "Puede de iniciar la depuración en el servidor Web" mensaje de error cuando ejecuta unas aplicaciones Web ASP.NET.
  Describe un problema que se produce cuando utiliza localhost como el servidor Web para crear una aplicación ASP.NET y utiliza la dirección IP para la Identificación del sitio Web. Para resolver el problema, modifique el archivo de proyecto .webinfo y...

Más Solución de problemas...

¿Necesita más ayuda?
Contar con un profesional de soporte técnico por correo electrónico, online o por teléfono
El control de usuario de ActiveX no está visible en el diseñador de formularios cuando se vuelve a abrir un proyecto existente

Advertencia: Artículo de Traducción Automática, vea la exención de responsabilidad.

Para solucionar este problema, quitar las existentes las referencias de control de usuario de ActiveX y, a continuación, agregue las referencias actualizadas de nuevo. Para ello, siga uno de los métodos siguientes.

Volver al principio

1 (Método)

1. En el Explorador de soluciones, bajo referencias, haga clic con el botón secundario del mouse en AxProject1 y, a continuación, haga clic en Quitar.
2. En referencias, haga clic con el botón secundario del mouse en Project1 y, a continuación, haga clic en Quitar.
3. En el cuadro de herramientas, haga doble clic en Project1.TestControl para agregar el control de usuario de ActiveX modificado en el formulario.
The ActiveX user control is not visible in the form designer when you reopen an existing project

Warning: Article on Machine Translation, see disclaimer.

To solve this problem, remove the existing references to user control of ActiveX, and then add references updated again. To do this, follow one of the following methods.

1 (Method)

1. In Solution Explorer, under References, right-click the secondary mouse AxProject1, and then click Remove.
2. References, click with the secondary mouse button in Project1, and then click Remove.
3. In the toolbox, double-click Project1.TestControl to add the user control of ActiveX in the amended form.

Note that the ActiveX user control in the Windows Forms designer and reference to Project1 and AxProject1 listed under references to UserControlDemo.
4. In the menu, click Build Solution.
The ActiveX user control is not visible in the Form Designer when you reopen an existing project

To work around this problem, remove the existing references of the ActiveX user control, and then add the updated references again. To do this, follow one of the following methods.

**Method 1**

1. In Solution Explorer, under References, right-click **AxProject1**, and then click **Remove**.
2. Under References, right-click **Project1**, and then click **Remove**.
3. In the toolbox, double-click **Project1.TestControl** to add the modified ActiveX user control to the form.

Notice that the ActiveX user control in the Windows Form Designer and the references to **AxProject1** and **Project1** appear under References for **UserControlDemo**.
4. On the Build menu, click **Build Solution**.

You do not receive any errors.
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?)