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)

BLEU Evaluation Metric

- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it's precision-oriented
  - BLEU formula
    
    \[
    \text{BLEU} = \exp \left( \sum_{n=1}^{4} \frac{1}{n} \log p_n \right)
    \]

  - \(p_n = \frac{\text{counts n-grams up to length 4}}{\max(\text{words-in-reference / words-in-machine} - 1, 0)}\)

  - \(p_1 = 1 \text{-gram precision}\)
  - \(p_2 = 2 \text{-gram precision}\)
  - \(p_3 = 3 \text{-gram precision}\)
  - \(p_4 = 4 \text{-gram precision}\)

- BLEU Evaluation Metric

  - 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 some portion of reference translation twice (can't cheat by typing out the same sentence twice)

  - 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….

Illustrative translation results

- "la politique de la haine.
  - (Foreign Original)
  - "the policy of the hatred.
  - (IBMT+N-grams+Stack)

- "nous avons signé le protocole.
  - (Foreign Original)
  - "we have signed the protocol.
  - (IBMT+N-grams+Stack)

- "où était le plan solide ?
  - (Foreign Original)
  - "but where was the solid plan?
  - (IBMT+N-grams+Stack)

- "where was the economic base?"

Bilingual Evaluation Understudy

- BLEU Evaluation Metric

  - Reference (human) translation: The U.S. island of Guam is maintaining a high state of alert after the Guam airport was shot by a gunman calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

  - Machine translation: The Atlantic (foreign language article) said the office of the foreign language article said the name of the Arab rich business [?] and so on.

  - The thread will be also after public place and then on the airport to start the biochemical attack [?] high state of alert for maintenance.

  - BLEU in Action

  - 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 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)
A complete translation system

Decoding for IBM Models

- Of all conceivable English word strings, find the one maximizing $P(e) \times P(f \mid e)$
- 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)

slide from G. Doddington (NIST)
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

- 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

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
\[ \arg\max_e P(e | f) = \]
\[ = \arg\max_e P(e) \times P(f | e) / P(f) \]
\[ = \arg\max_e P(e) \times P(f | e) \times \text{length}(e)^{1.1} \times K \]

Rewards longer hypotheses, since these are 'unfairly' punished by \( P(e) \)

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.)
Phrase-Based Statistical MT

• Foreign input segmented into phrases
  – “phrase” is any sequence of words
• Each phrase is probabilistically translated into English
  – \( P(\text{to the conference} | \text{zur Konferenz}) \)
  – \( P(\text{into the meeting} | \text{zur Konferenz}) \)
• 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” \( \rightarrow \) …
  – “interest in” \( \rightarrow \) …
• 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.
How to Learn the Phrase Translation Table?

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

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

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) => Translation model => Sentence (Japanese)
  - Preprocess English by a parser
  - Probabilistic Operations on a parse-tree
    1. Record child nodes
    2. Insert extra nodes
    3. Translate leaf words
Parse Tree (E) → Sentence (J)

Parameter Table: Reorder

Parameter Table: Insert

3. Translate

Conditioning Feature = word (E) identity

Conditioning Feature = Child label Sequence

Conditioning Feature = Parent Label & Node Label (position)

none (word selection)
Parameter Table: Translate

<table>
<thead>
<tr>
<th>E</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>daizuki</td>
<td>1.00</td>
<td>kara</td>
<td>0.333</td>
<td>wo</td>
</tr>
<tr>
<td></td>
<td>kara</td>
<td>0.333</td>
<td>ni</td>
<td>0.000</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td>kara</td>
<td>0.333</td>
<td>ni</td>
<td>0.000</td>
<td>no</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: Processed by Collins Parser [Collins, 1999]
  - Trained: 40K Treebank, Accuracy ~90%
- E: Flattened parse tree
  - To capture word-order difference (SVO->SOV)
- EM Training: 20 Iterations
  - 50 minutes (Sparc 200MHz 1-CPU) or
  - 30 seconds (Pentium 3 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

```
He aimed a revolver at me

彼は拳銃を私に向けた

Syntax-based model
IBM Model 3
```

Result: Alignment 3

```
He has unusual ability in English

彼は英語にすぐ能を持てている

Syntax-based Model
IBM Model 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 for the future
  - Statistical phrases
  - Syntax-based models
  - Better language models in other respects (e.g., grammar)
  - Usably efficient decoding models (by restricting model?)