Multiple Reference Translations

Reference translation 1:

The USA International Airport of Guam and other public places. Guam Authority has been on alert.

Reference translation 2:

The USA International Airport of Guam and other public places. Guam Authority has been on alert.

Reference translation 3:

Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and that threatened to launch a biochemical attack on such public places as the airport.

Reference translation 4:

Guam airport and its offices both received an e-mail from someone claiming to be a self-claimed Arabian millionaire and threatening a biological/chemical attack against public places such as the airport.

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.
- Nous avons signé le protocole.
- Où était le plan solide?

Machine translation:

- 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)

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?
  - 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)

*** Amazingly 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)
  \[
  \exp \left( 1.5 \log p_1 + 0.5 \log p_2 + 0.25 \log p_3 + 0.125 \log p_4 \right)
  \]
  \[
  p_1 = \text{1-gram precision}
  p_2 = \text{2-gram precision}
  p_3 = \text{3-gram precision}
  p_4 = \text{4-gram precision}
  \]
  \[
  \text{max(words-in-reference} / \text{words-in-machine} - 1, 0)
  \]
BLEU Tends to Predict Human Judgments

<table>
<thead>
<tr>
<th>Human Judgments</th>
<th>NIST Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.5</td>
<td>-2.0</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.5</td>
</tr>
<tr>
<td>-1.5</td>
<td>-1.0</td>
</tr>
<tr>
<td>-1.0</td>
<td>-0.5</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>1.0</td>
<td>1.5</td>
</tr>
<tr>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>2.0</td>
<td>2.5</td>
</tr>
</tbody>
</table>

**BLEU in Action**

<table>
<thead>
<tr>
<th>(Foreign Original)</th>
<th>(Reference Translation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>the gunman was shot to death by the police.</td>
<td>the gunman was police kill. #1</td>
</tr>
<tr>
<td>the gunman was police kill.</td>
<td>wounded police jaya of #2</td>
</tr>
<tr>
<td>the gunman was shot dead by the police.</td>
<td>the gunman arrested by police kill . #3</td>
</tr>
<tr>
<td>the gunman arrested by police kill.</td>
<td>the gunmen were killed. #4</td>
</tr>
<tr>
<td>the gunmen were killed.</td>
<td>the gunman was shot to death by the police . #5</td>
</tr>
<tr>
<td>the gunmen were killed by police ?SUB&gt;0 ?SUB&gt;0</td>
<td>al by the police . #6</td>
</tr>
<tr>
<td>al by the police .</td>
<td>the ringer is killed by the police . #7</td>
</tr>
<tr>
<td>the ringer is killed by the police .</td>
<td>police killed the gunman . #8</td>
</tr>
<tr>
<td>police killed the gunman.</td>
<td>(good)</td>
</tr>
</tbody>
</table>

**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**

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)
\[
\hat{e} = \arg\max_{e} P(e | f) = \\
\arg\max_{e} P(e) \times P(f | e) / P(f) = \\
\arg\max_{e} P(e) \times P(f | e) = e
\]

What StatMT people do in the privacy of their own homes
\[
\arg\max_{e} P(e | f) = \\
\arg\max_{e} P(e) \times P(f | e) / P(f) = \\
\arg\max_{e} P(e)^{2.4} \times P(f | e) \times \text{length}(e)^{1.1} \times KS^{3.7} \ldots \text{works better!}
\]

Which model are you now paying more attention to?

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(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” \(\rightarrow\) …
  - “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.

Mary did not slap the green witch

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.

IBM Models are 1-to-Many

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

Mary did not slap the green witch

How to Learn the Phrase Translation Table?

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

Mary did not slap the green witch

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

- 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)

Parameter Table: Reorder

| Original Order | Reordering | P(reorder|original) |
|----------------|------------|---------------|
| PRP VB1 VB2    | PRP VB1 VB2 | 0.074         |
| PRP VB2 VB5    | VB1 VB1 VB2 | 0.061         |
| VB1 PRP VB2    | VB1 VB1 VB2 | 0.037         |
| VB2 PRP VB1    | VB1 VB1 VB2 | 0.083         |
| VB2 VB1 VB2    | VB1 VB1 VB2 | 0.021         |
| VB TO VB       | VB TO VB    | 0.693         |
| VB TO NN       | VB TO NN    | 0.251         |
| TO NN NN       | NN TO NN    | 0.749         |

Parameter Table: Insert

Parameter Table: Translate
Parameter Table: Translate

<table>
<thead>
<tr>
<th>E</th>
<th>Jdaisuki 1.000</th>
<th>NULL</th>
<th>0.016</th>
<th>0.003</th>
<th>0.003</th>
<th>0.333</th>
<th>0.333</th>
<th>0.333</th>
<th>0.333</th>
</tr>
</thead>
<tbody>
<tr>
<td>kana</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>kanb</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>null</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>555</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</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

Ave. Score   # perf sent
Y/K Model     0.582     10
IBM Model 5 0.431     0

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

Result: Alignment 1

Synth-based Model
He adores listening to music
彼 は 音楽 を 聞く の が 大好き です

IBM Model 2
He adores listening to music
彼 は 音楽 を 聞く の が 大好き です

Result: Alignment 2

Synth-based Model
彼 は 拳銃 を 私 に 向け た

IBM Model 3
He aimed a revolver at me
彼 は 拳銃 を 私 に 向け た

Result: Alignment 3

Synth-based Model
He has unusual ability in English
彼 は 英語 に すばら な 能 力 を 持つ て いる

IBM Model 3
He has unusual ability in English
彼 は 英語 に すばら な 能 力 を 持つ 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

Expand by Model

Decoding Grammar for C/E 3M corpus

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, Googles) 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?)