Agenda for today

- Introduction to Machine Translation
  - Data-driven statistical machine translation
  - Translation models
    * Parallel corpora
    * Document-, sentence-, word-alignment
    * Phrase-based translation
  - MT decoding algorithm
  - Language models
  - MT evaluation
  - Further topics for exploration
Machine Translation

• Mapping from a source language string to a target language string, e.g.,

  Spanish source:
  Perros pequeños tienen miedo de mi hermanita torpe
  English target:
  Small dogs fear my clumsy little sister

• The “right way” to do this

  – Map the source language to some semantic interlingua, e.g.,
    fear(dog([plural],[small]),sister([my,singular],[young,clumsy]))
  – Generate the target string from the interlingual representation

• This isn’t feasible in current state of technology
Current best approaches to MT

• Statistical models are the current best practice
  – e.g., Google translation is data driven

• Basic approach taken from statistical speech recognition
  – Let source string be $f$ and target language be $e$

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

  – $P(f \mid e)$ is the translation model
    (akin to acoustic model in statistical speech recognition)

  – $P(e)$ is the language model
MT system

Summary

Pre–process

Translate

Post–process

Translation

Parallel Corpora

Phrase Table

Weights

Train Weights

Language Model

Train LM

Target Language Corpora

Bleu Evaluation

Reference(s)

Pre–process

Extract Phrases

Align Sentences

Align Words
Translation model

• Given a pair of strings $<f, e>$, assigns $P(f \mid e)$
  – If $f$ looks like a good translation of $e$, then $P(f \mid e)$ will be high
  – If $f$ doesn’t look like a good translation of $e$, then $P(f \mid e)$ will be low

• Where do these pairs of strings $<f, e>$ come from?
  – Paying people to translate from multiple languages is expensive
  – Would rather get free resources, even if imperfect (or “noisy”) data
  – Such data is produced independently: parallel corpora
Parallel corpora

- Examples:
  - The Hansards corpus of Canadian Parliament transcripts, by law in both French and English
  - Similar resources for EU official proceedings and documents
  - Software manuals, web pages, other available data

- Document-aligned

- Must be sentence- and word-aligned to derive models
Learning alignment models

• If we only have document-aligned parallel corpora, how do we get to the sentence alignment?

• Simple heuristics based on length of sentences.

• Once we have sentence-aligned parallel corpora, how do we get to the word alignment?

• One answer: align words that often appear together
Small dogs fear my clumsy little sister. Because she is so clumsy, the dogs think she will fall on them. Big dogs do not fear her, just the small ones. They do not fear my little sister because she fears them.

Perros pequeños tienen miedo de mi hermanita torpe. Porque es tan torpe, los perros creen que ella se caerá sobre ellos. Perros grandes no tienen miedo de ella, solo los pequeños. No tienen miedo de mi hermanita porque ella tiene miedo de ellos.
## Example sentence alignment

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small dogs fear my clumsy little sister</td>
<td>Perros pequeños tienen miedo de mi hermanita torpe</td>
</tr>
<tr>
<td>Because she is so clumsy, the dogs think she will fall on them</td>
<td>Porque es tan torpe, los perros creen que ella se caerá sobre ellos</td>
</tr>
<tr>
<td>Big dogs do not fear her, just the small ones</td>
<td>Perros grandes no tienen miedo de ella, solo los pequeños</td>
</tr>
<tr>
<td>They do not fear my little sister because she fears them</td>
<td>No tienen miedo de mi hermanita porque ella tiene miedo de ellos</td>
</tr>
</tbody>
</table>
Example word alignment

Perros pequeños tienen miedo de mi hermanita torpe

Small dogs fear my clumsy little sister
Example word alignment

Perros  pequeños  tienen  miedo  de  mi  hermanita  torpe

Small  dogs  fear  my  clumsy  little  sister
Notation

• Source string: $f = f_1 \ldots f_{|f|}$
• Target string: $e = e_1 \ldots e_{|e|}$
• Alignment under the assumption of at most one target word per source word: $a = a_1 \ldots a_{|f|}$, where $0 \leq a_i \leq |e|$
• $a_i = j$ if $f_i$ aligns with $e_j$
• $a_i = 0$ if $f_i$ is unaligned with anything in $e$
• Thus for our example:

$$f = \text{Perros pequeños tienen miedo de mi hermanita torpe}$$
$$e = \text{Small dogs fear my clumsy little sister}$$
$$a = 2 \ 1 \ 3 \ 3 \ 0 \ 4 \ 7 \ 5$$
Probabilistic modeling

- Given a target string, assign joint probabilities to source strings and alignments: $P(f, a | e)$
- The probability of the source string is the sum over all alignments
  $$P(f | e) = \sum_a P(f, a | e)$$
- The best alignment is the one that maximizes the probability
  $$\hat{a} = \arg\max_a P(f, a | e)$$
- Decompose full joint into product of conditionals:
  $$P(f, a | e) = P(F | e) \prod_{i=1}^{F} P(f_i, a_i | e f_1 a_1 \ldots f_{i-1} a_{i-1})$$
  where $F = |f|$
Heuristic alignments

- Calculate word similarity in some way, e.g., Dice coefficient

\[
dice(i, j) = \frac{2c(e_i, f_j)}{c(e_i) c(f_j)}
\]

where \( c(e_i, f_j) \) is the count of parallel sentences containing \( e_i \) on the source side and \( f_j \) on the target side

- Build matrix of similarities

- Align highly-similar words

- Various strategies to align:
  - Choose \( a_j = \argmax_i \{dice(i, j)\} \)
  - Greedily choose best link (globally), then remove row and column from matrix (competitive linking algorithm)
Alignment algorithms

- Heuristic
  - Dice
  - Competitive linking

- Statistical
  - IBM models 1-5 [Brown et al. 93]
    * Expectation-Maximization algorithm
    * Another pipeline
  - HMM model [Deng & Byrne 05]
  - GIZA++ software [code.google.com/p/giza-pp/]

15
Limitations of word-based translation

- One-to-many and many-to-many alignment
  - Some approaches make simplifying assumptions regarding word “fertility”, i.e., number of aligned words
- Crossing alignments
  - Relatively small permutations
    * e.g., post-nominal modifiers (perros pequeños ⇒ small dogs)
  - Relatively large permutations
    * e.g., argument ordering (‘in pain young Skywalker is’)
Example word alignment

Perros pequeños tienen miedo de mi hermanita torpe

Small dogs fear my clumsy little sister
Phrase-based translation

• Translate sequences of source-language words into (possibly) sequences of target-language words

• Advantages of phrase-based translation
  – Many-to-many translation
  – Allows for more context in translation

• Phrase table
  – Extracted by “growing” word alignments
  – Limited by phrase length
  – Ambiguity in translation look-up
Extracting phrases from word-alignments

<table>
<thead>
<tr>
<th>Maria no daba una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
</table>

Mary

did

not

slap

the

green

witch
Extracting phrases from word-alignments
Extracting phrases from word-alignments

Maria no daba una bofetada a la bruja verde

Mary did not slap the green witch
Extracting phrases from word-alignments
Decoding algorithm

- Moses decoder [www.statmt.org/moses/]
  - Beam search
  - Build English (target language sentence) by hypothesis expansion (left-to-right)
  - Ambiguity
  - Search space pruning
MT system
Language model

• Goal: to detect “good” English

• Standard technique: $n$-gram model
  – Calculate the probability of seeing a sequence of $n$ words
  – Probability of a sentence is product of $n$-gram probabilities

• Bi-gram model example:

  $P(\text{Small dogs fear my clumsy little sister}) =$

  $P(\text{Small}) \times P(\text{dogs}|\text{Small}) \times P(\text{fear}|\text{dogs}) \times P(\text{my}|\text{fear}) \times$

  $P(\text{clumsy}|\text{my}) \times P(\text{little}|\text{clumsy}) \times P(\text{sister}|\text{little})$

• Arbitrary values of $n$
  – Language modeling, v0.0: $n=2$
Estimating language model from corpora

- Probabilities estimated via maximum likelihood
  \[
P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}
  \]
  e.g.:
  \[
P(\text{dog}|\text{Small}) = \frac{C(\text{Small dog})}{C(\text{Small})}
  \]
- Unobserved \(n\)-grams get zero probability!
- Smoothing to reserve probability mass for unobserved events
- Corpus size matters
  - Language modeling corpus, v0.0: 40k sentences
MT system

Summary → Pre-process → Translate → Post-process → Translation

Phrase Table → Weights

Parallel Corpora → Pre-process → Extract Phrases

Pre-process → Align Sentences

Extract Phrases → Align Words

Train Weights

Language Model → Train LM

Train LM → Target Language Corpora

Reference(s) → Bleu Evaluation

Bleu Evaluation → Reference(s)
MT evaluation

• Ideal: human evaluation
  – Adequacy: does the translation correctly capture the information of the source sentence?
  – Fluency: is the translation a “good” sentence of the target language?
  – But: slow and expensive

• Automatic evaluation
  – Intuition: comparing two candidate translations $T_1$ and $T_2$
    * To the extent that $T_1$ overlaps more with a reference (human-produced) translation $R$, it is “better” than $T_2$
  – How to measure overlap?
  – Differences in length of translation?
  – Multiple reference translations?
BLEU

• Measure overlap by counting \( n \)-grams in candidate that match the reference translation

• More matches \( \Rightarrow \) better translation

• Precision metric

• Brevity penalty

\[
\log \text{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log(p_n)
\]
Brief note on text processing

- Tokenization
- Casing
Further topics of exploration

- Translation model
  - More, better, different data
  - Different word-alignment algorithms
  - Length of extracted phrases

- Language model
  - More, better, different data
  - Size of $n$-grams

- Add more knowledge to the process
  - Numbers
  - Dates
  - Named entities