Agenda

- 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

$$\begin{aligned}\operatorname*{argmax}_e \mathrm{P}(e \mid f) &= \operatorname*{argmax}_e \frac{\mathrm{P}(f \mid e)\,\mathrm{P}(e)}{\mathrm{P}(f)} \\ &= \operatorname*{argmax}_e \mathrm{P}(f \mid e)\,\mathrm{P}(e)\end{aligned}$$

  - $\mathrm{P}(f \mid e)$ is the translation model (akin to acoustic model in statistical speech recognition)
  - $\mathrm{P}(e)$ is the language model

Translation model

- Given a pair of strings $<f, e>$, assigns $\mathrm{P}(f \mid e)$
  - If f looks like a good translation of e, then $\mathrm{P}(f \mid e)$ will be high
  - If f doesn't look like a good translation of e, then $\mathrm{P}(f \mid e)$ 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

Example parallel corpus

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

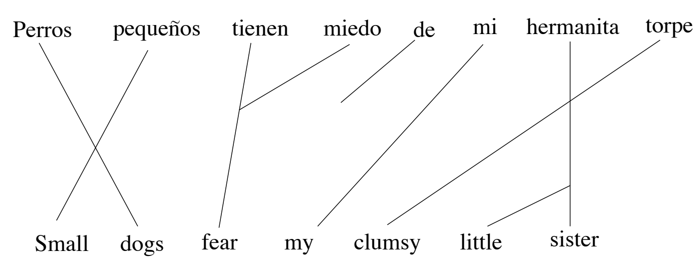
Small dogs fear my clumsy little sister	Perros pequeños tienen miedo de mi hermanita torpe
Because she is so clumsy, the dogs think she will fall on them	Porque es tan torpe, los perros creen que ella se caerá sobre ellos
Big dogs do not fear her, just the small ones	Perros grandes no tienen miedo de ella, solo los pequeños
They do not fear my little sister because she fears them	No tienen miedo de mi hermanita porque ella tiene miedo de ellos

Example word alignment

Perros pequeños tienen miedo de mi hermanita torpe

Small dogs fear my clumsy little sister

Example word alignment



Notation

- Source string: $\boldsymbol{f} = \boldsymbol{f}_1 \ldots \boldsymbol{f}_{|\boldsymbol{f}|}$
- Target string: $\boldsymbol{e} = \boldsymbol{e}_1 \ldots \boldsymbol{e}_{|\boldsymbol{e}|}$
- Alignment under the assumption of at most one target word per source word: $\boldsymbol{a} = \boldsymbol{a}_1 \ldots \boldsymbol{a}_{|\boldsymbol{f}|}$, where $0 \leq \boldsymbol{a}_i \leq |\boldsymbol{e}|$
- $\boldsymbol{a}_i = \boldsymbol{j}$ if $\boldsymbol{f}_i$ aligns with $\boldsymbol{e}_j$
- $\boldsymbol{a}_i = 0$ if $\boldsymbol{f}_i$ is unaligned with anything in $\boldsymbol{e}$
- Thus for our example:

  $\boldsymbol{f}$ = Perros pequeños tienen miedo de mi hermanita torpe

  $\boldsymbol{e}$ = Small dogs fear my clumsy little sister

  $\boldsymbol{a}$ = 2 1 3 3 0 4 7 5

Probabilistic modeling

- Given a target string, assign joint probabilities to source strings and alignments: $\mathrm{P}(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$
- The probability of the source string is the sum over all alignments

$$\mathrm{P}(\boldsymbol{f} \mid \boldsymbol{e}) = \sum_{\boldsymbol{a}} \mathrm{P}(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$$

- The best alignment is the one that maximizes the probability

$$\hat{\boldsymbol{a}} = \operatorname*{argmax}_{\boldsymbol{a}} \mathrm{P}(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e})$$

- Decompose full joint into product of conditionals:

$$\mathrm{P}(\boldsymbol{f}, \boldsymbol{a} \mid \boldsymbol{e}) = \mathrm{P}(\boldsymbol{F} \mid \boldsymbol{e}) \prod_{i=1}^{\boldsymbol{F}} \mathrm{P}(\boldsymbol{f}_i, \boldsymbol{a}_i \mid \boldsymbol{e} \boldsymbol{f}_1 \boldsymbol{a}_1 \ldots \boldsymbol{f}_{i-1} \boldsymbol{a}_{i-1})$$

where $\boldsymbol{F} = |\boldsymbol{f}|$

Heuristic alignments

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

$$\text{dice}(i, j) = \frac{2c(e_i, f_j)}{c(e_i)c(f_j)}$$

  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 $a_j = \text{argmax}_i\{\text{dice}(i, j)\}$
  - Greedily choose best link (globally), then remove row and column from matrix (*competitive linking* algorithm)

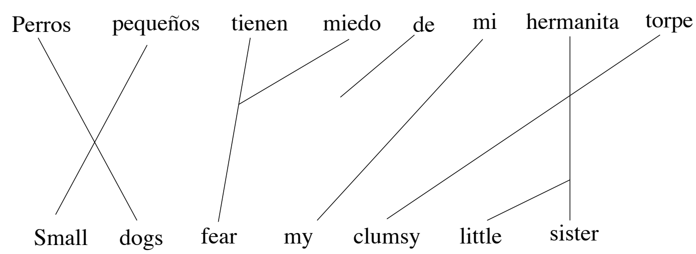
Alignment algorithms

- **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/]**

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 $\Rightarrow$ small dogs)
  - Relatively large permutations
    - e.g., argument ordering ('in pain young Skywalker is')

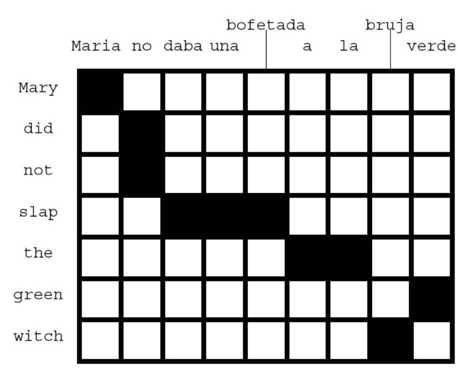
Example word alignment



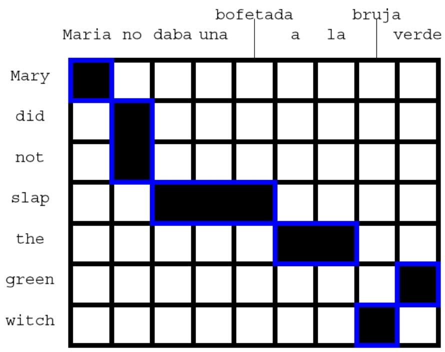
Phrase-based translation

- **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**

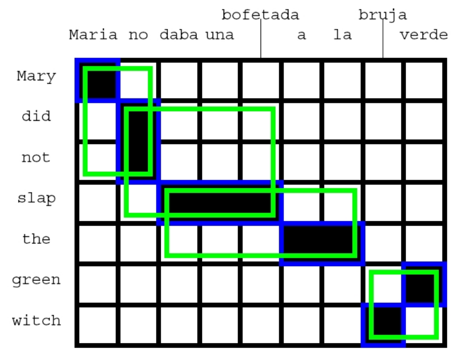
Extracting phrases from word-alignments



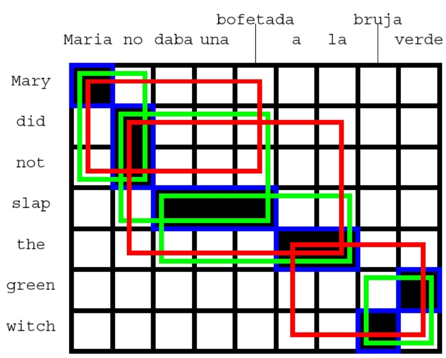
Extracting phrases from word-alignments



Extracting phrases from word-alignments



Extracting phrases from word-alignments



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:

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

  $\mathrm{P}(Small) * \mathrm{P}(dogs|Small) * \mathrm{P}(fear|dogs) * \mathrm{P}(my|fear)*$

  $\mathrm{P}(clumsy|my) * \mathrm{P}(little|clumsy) * \mathrm{P}(sister|little)$
- Arbitrary values of n
  - Language modeling, v0.0: n=2

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

$$\log \mathrm{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log(p_n)$$

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