Natural Language Processing
CSE 325/425

Sihong Xie

Lecture 25:
- Decoding of machine translation
- Neural translation via seq2seq models

Decoding

- We have describe two alignment models
  - IBM Model-1
  - HMM $$P(F, A|E) = P(J|I) \times \prod_{j=1}^{J} P(a_j|a_{j-1}, I) P(f_j|e_{a_j})$$

model parameters (from EM)

- How to find the best translation *E* given a foreign sentence F and the model:

$$\hat{E} = \underbrace{\operatorname*{argmax}_{E \in \text{English}}}_{\text{unknown}} \overbrace{P(F|E)}^{\text{translation model}} \overbrace{P(E)}^{\text{language model}}$$

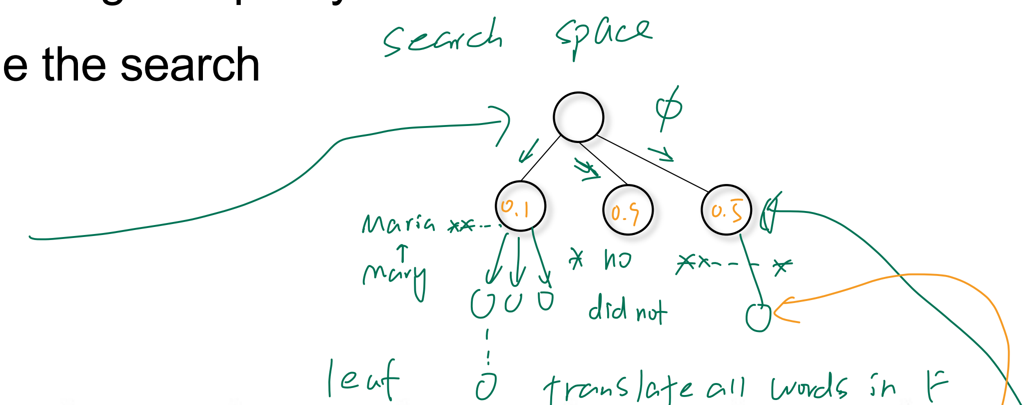
given

  - Note that E is unknown and it is different from aligning *E* and *F.*
  - Decoding with a bigram language model is NP-complete (Knight 1999).

Knight, K. (1999). Decoding complexity in word-replacement translation models. *Computational Linguistics*, *25*(4), 607– 615.

Decoding

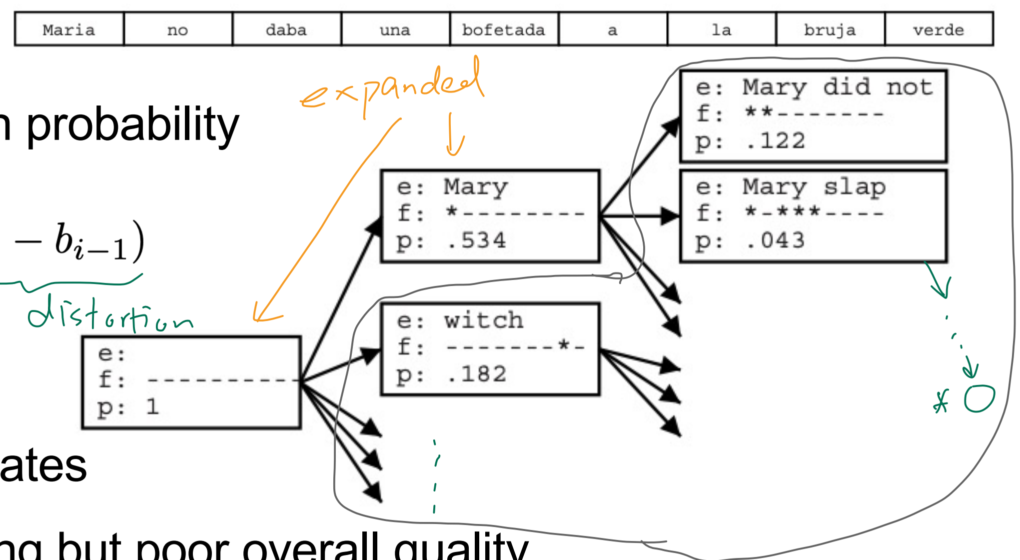
- Need greedy heuristic search algorithms.
  - Believe that greedy is sufficient to find good quality translations.
  - Use heuristic cost function to guide the search
- Best-first search
  - state (node): partial translations



Decoding

- Best-first search
  - state (node): partial translations
  - expansion (edge): translating one more word/phrase from previous states
  - expand the "best" state
    - measured in partial translation probability

$$P(E_{\text{Partial}}) = \prod_{i \in \text{Partial}} \underbrace{\phi(\bar{f}_i, \bar{e}_i)}_{\text{phrase translation table}} \underbrace{d(a_i - b_{i-1})}_{\text{distortion}}$$



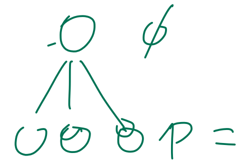
    - expensive: maintains many states

      local maximum: good beginning but poor overall quality

when stop with a leaf node with the highest prob compared to all states in the priority queue.

priority queue

frontier after the second expansion



$g(p) = \phi(\text{Maria} \mid \text{Mary})$

$p =$ Maria / Mary — — — — — Remaining

$e_2^*\ e_3^* \ldots e_6^*$

Decoding

- A* search (AI course)
  - Use a function to evaluate both **current** and **future** quality.

HMM/model-1 $\neq$ $h^*(p) = \prod_{i=2}^{6} \phi(\bar{f}_i, \bar{e}_i^*)$

non-negative

$$f^*(p) = g(p) + h^*(p)$$

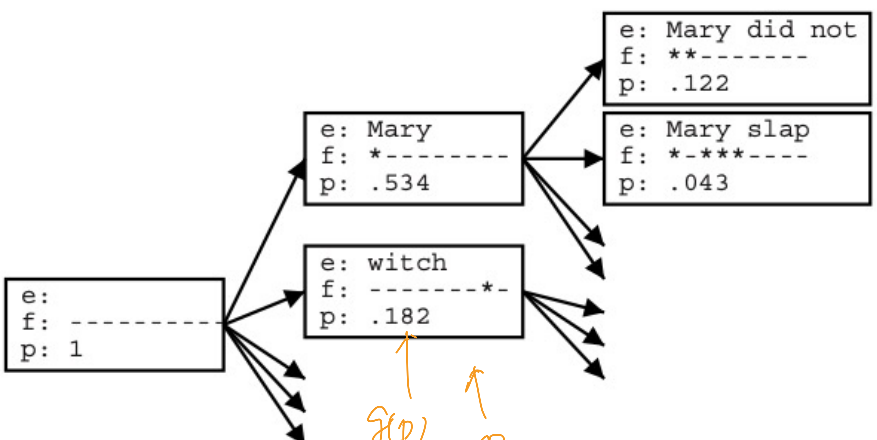
The quality of the partial translation

$P(E_{partial})$

best possible translation quality for the yet-to-translate foreign words: expensive to estimate and needs heuristic.

$h^*(p) = \prod_{i \in Rem} \phi(\bar{f}_i, \bar{e}_i^*)$

Maria	no	daba	una	bofetada	a	la	bruja	verde

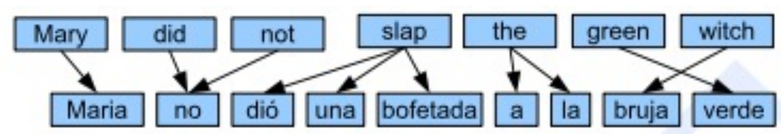


$g(p) + h^*(p) = f^*(p)$

p

best translation

An unknown good translation for your reference

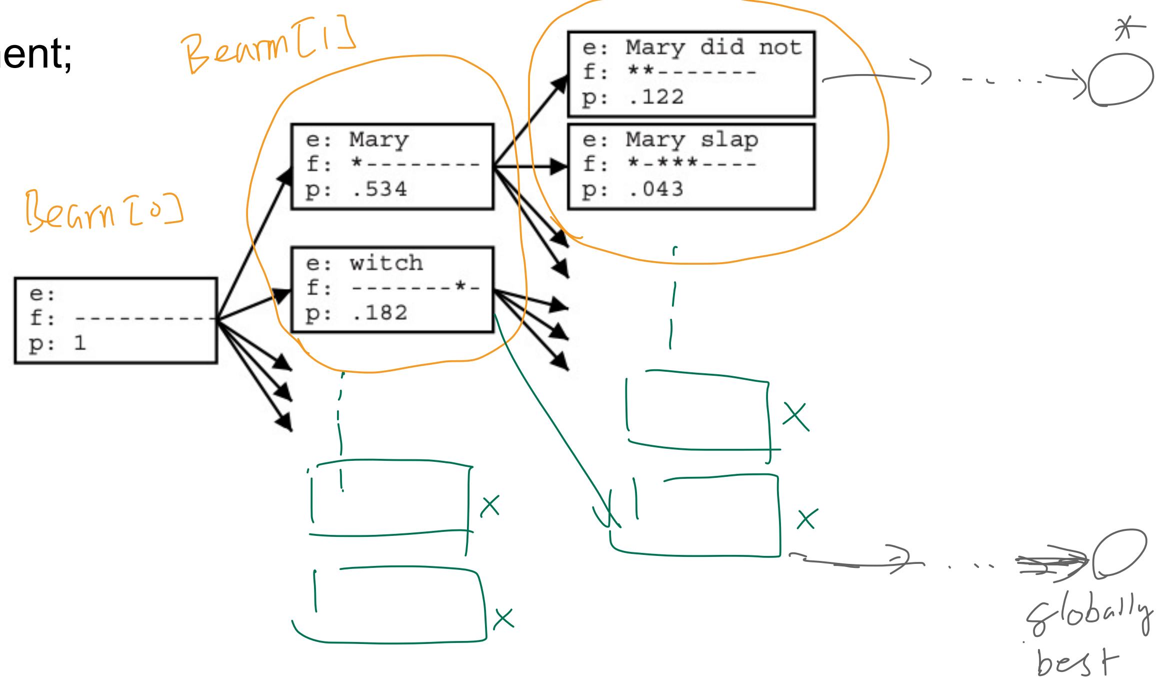


Decoding

note:
widely used in NLP or computer vision

- Beam search (combined with best-first or A* searches)
  - Beam = a small set of the current best states.
  - Expand those states in the beam and only keep the best expansions.

    cut down the # of states that have to be kept in memory.
  - Example: beam size = 2
    - after the expansion, only keep the two best states
    - after the second expansion, again only keep the best two states.
    - less memory size requirement;
    - can have local minima;
    - combined with A* heuristic.



Decoding

ϕ $P(\text{Partial}) = 1$

Maria → Mary $P(\quad) = 0.534$

xx x - - - x $P(\quad) = 0.00001$

- Multi-stack search → priority queue
  - Translation of different number of foreign words can't be compared directly.

$$P(E_{\text{Partial}}) = \prod_{i \in \text{Partial}} \phi(\bar{f}_i, \bar{e}_i) d(a_i - b_{i-1})$$ in best-first search

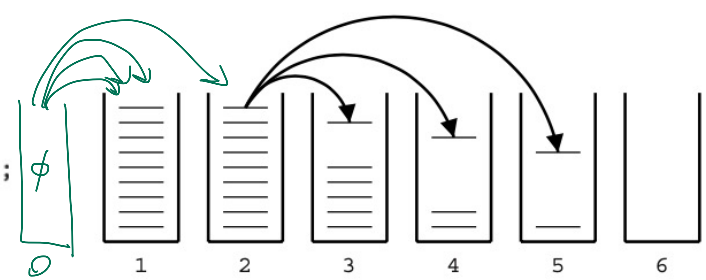
  - The more words/phases translated, the smaller the probability.
  - Need a stack (priority queue) for each number of foreign words translated.

nf : # of foreign phases in F

hypothesis ⇔ partial translation

```
initialize hypothesisStack[0 .. nf];
create initial hypothesis hyp_init;
add to stack hypothesisStack[0];
for i=0 to nf-1:
  for each hyp in hypothesisStack[i]:
    for each new_hyp that can be derived from hyp:
      nf[new_hyp] = number of foreign words covered by new_hyp;
      add new_hyp to hypothesisStack[nf[new_hyp]];
      prune hypothesisStack[nf[new_hyp]];
find best hypothesis best_hyp in hypothesisStack[nf];
output best path that leads to best_hyp;
```

pointing to the dst stack



partial translations covering 1 Foreign pharse

covering 2

covering 6

Evaluation

- Human evaluation faithfulness fluency - - - expensive { ① hire humans ② the eval can't be reused. !!!
- Automatic evaluation using BLEU score
  - Intuition: compute how often a predicted translation matches n-grams in any of the ground-truth translations. ← spend $ & time but can be reused !!! LM

unigram precision:
17/18
8/14

Cand 1: It is a guide to action which ensures that the military always obeys the commands of the party

Cand 2: It is to insure the troops forever hearing the activity guidebook that party direct

F

Ref 1: It is a guide to action that ensures that the military will forever heed Party commands

Ref 2: It is the guiding principle which guarantees the military forces always being under the command of the Party

Ref 3: It is the practical guide for the army always to heed the directions of the party

multiple correct & good translations

BLEU: a method for automatic evaluation of machine translation, ACL, 2002

[Caution of using BLEU: https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/](https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/)

bi-gram precision ?

HW 6

Evaluation

- Compute precision for *n*-grams, for *n=1,2,3,4* for all candidate translations predicted by the model.

candidate translatio

$$\text{prec}_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} \text{Count}_{\text{match}}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} \text{Count}(n\text{-}gram')}$$

  - Then take the geometric mean of these precisions $\left(\prod_{n=1}^{4} \text{prec}_n\right)^{\frac{1}{4}}$ as the BLEU score of translating a sentence

BLEU: a method for automatic evaluation of machine translation, ACL, 2002

[Caution of using BLEU: https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/](https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/)

Evaluation

- Pitfalls in the evaluation using BLEU
  - Very short translations can have high precisions.
  - Example: "*for the*" will have precision 1 when the references contains the "*for the*", regardless of the remaining words.

BLEU = penalty of brevity $\times \left(\prod_n \mathrm{Prec}_n\right)^{1/4}$

Ref 3: It is the practical guide for the army always to heed the directions of the party

  - Repetitive matches
    - Translation: "the the the the the the the"
    - Reference: "the cat is on the mat"
    - Modified precision should be just 2/7 (2 = max number of "*the*" in the reference)

$\text{precision}_{n=1} = 1 = 7/7$

BLEU: a method for automatic evaluation of machine translation, ACL, 2002
Caution of using BLEU: https://slator.com/technology/how-bleu-measures-translation-and-why-it-matters/