Machine Translation

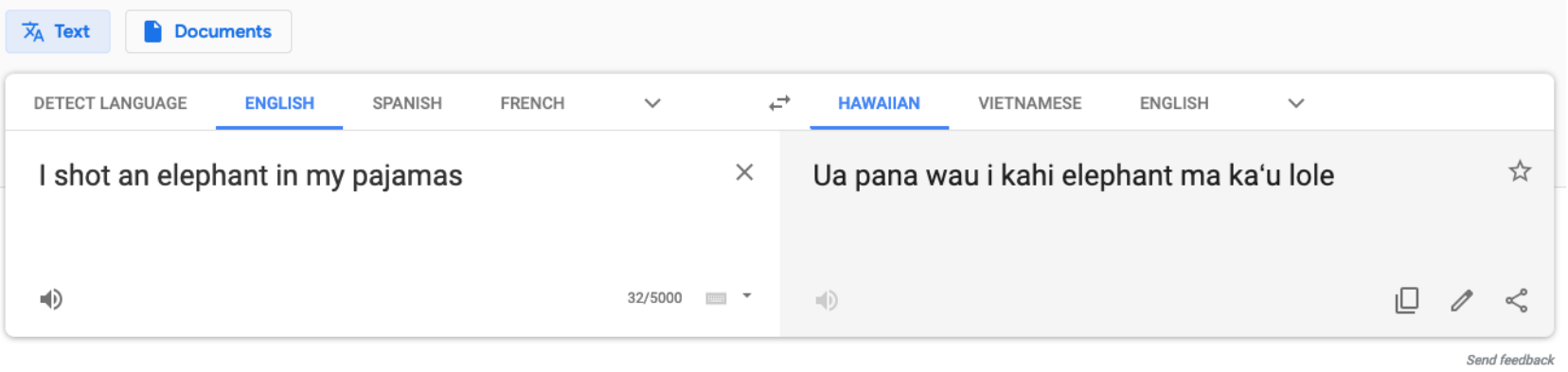
Bonan Min

bonanmin@gmail.com

Some slides are based on class materials from Thien Huu Nguyen, Yifan He, Dan Jurasky, Raymond Mooney, and others

Machine Translation (MT)

https://translate.google.com/



Machine Translation (MT)

MT Objectives

Rough translation

- for end users: web browsing / for computers: cross-lingual IE

Computer-aided human translation

- for formal documents; helpful when human post-editing is faster than full human translation

Fully automatic high-quality translation (FAHQT)

- feasible only for narrow sublanguages with restricted semantics
- early example: METEO (French-English translation of weather forecasts)

Terminologies

Source language F (foreign, e.g., French)

Target language E (English)

Parallel corpus

- The same document in two (or more) languages
- Translation is expected to be faithful on sentence level
- Comparable (English and French Wiki articles on Napoleon) vs. Parallel (English and French versions of a speech at the European Parliament)

Obstacles To Good MT

Lexical translation problems

Reordering problems

- Local reordering
- Long-distance reordering

Obstacles To Good MT

Lexical translation problems

- lexical divergences
  - need to disambiguate word senses in order to translate, e.g., for homonymous (e.g., bass) or polysemous (e.g., know below) words
    - E.g., knowing a fact or proposition (i.e., I know that snow is white) vs. familiarity with a person or location (i.e., I know John) -> French: savoir vs connaître
  - personal pronoun in some language does not distinguish gender (English does)
  - French requires specifying adjective gender (English doesn't)

- pro-drop (pronouns can be omitted in some languages)
  - must identify and resolve implied arguments when translating into English

- argument marking
  - English marks semantic roles by position; Japanese by postpositions; Russian by inflection
    - English: He adores listening to music
    - Japanese: kare ha ongaku wo kiku no ga daisuki desu
      he **music** **to** **listening** **adores**

Obstacles To Good MT: Reordering

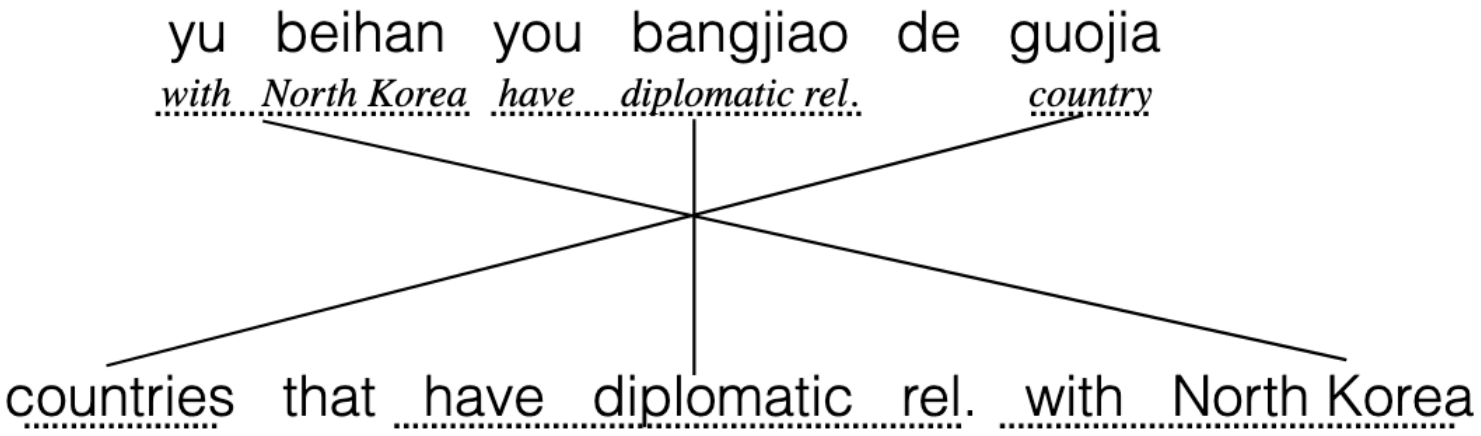
Local reordering:

- English adjectives precede nouns; French and Spanish adjectives normally follow nouns
- Argument structure and linking: Head-marking languages (marking the relation between the head and its dependents on the head) vs. dependent-marking languages (marking the relation on the non-head)
  - English: the man^{-A}'s ^{H} house
  - Hungarian: az ember ^{H} ház^{-A} a
    the man house-his

Long distance-reordering

- English, German, French, Mandarin are SVO, Japanese and Hindi are SOV
- Relative clauses are after the head in English, but often before the head noun in Mandarin

Obstacles To Good MT: Reordering



Rule-based MT

Direct MT: word-for-word + local re-ordering (need large bilingual vocabulary and simple reordering rules (e.g., moving adjectives after nouns when translating from English to French)

Transfer MT: incorporates parsing and parse-tree reordering (need distinct transfer rules for each pair of languages)

Interlingua: universal semantic representation (analyze the source language text into some abstract meaning representation, then generate target language from this interlingual representation)

Systran: started development in 1968; now hybrid rule-based/statistical system

Statistical MT

Relies on sentence-aligned bi-text

Canadian Hansards (French-English); Hong Kong Hansards (Chinese - English); European Parliament Proceedings (21 European languages)

Nowadays mine the Web for bitexts

Automatic sentence alignment (based on length, known word translations or word alignment)

Statistical MT

Data-driven approach: making use of available parallel corpora

Rule-based MT focuses on the process, statistical MT focuses on the result.

- It is impossible for a sentence in one language to be a translation of a sentence in another, strictly speaking (e.g., one cannot really translate Hebrew "*adonai roi*" (*the Lord is my shepherd*) into the language of a culture that has no sheep).
- The trade-off:
  - Clear in target language, but cost the fidelity to the original: *the Lord will look after me* (no sheep included)
  - Faithful to the original, but be obscure to the target language: *the Lord is for me like somebody who looks after animals with cotton-like hair* (not natural in English).
  - So, being both faithful to the source language and natural as an utterance in the target language is sometimes impossible. Professional translators actually need to balance between these two criteria in practice.
  - Statistical MT achieves these via the noisy channel model

Statistical MT: The Noisy Channel Model

Suppose we are translating French/Foreign sentences $\boldsymbol{F}$ to English $\boldsymbol{E}$

Using Bayes's rules:

$$\begin{aligned}\text{argmax}_E\, p(E|F) &= \text{argmax}_E \frac{p(F|E)\, p(E)}{p(F)} \\ &= \text{argmax}_E \boxed{p(F|E)}\,\boxed{p(E)}\end{aligned}$$

translation model language model

Statistical MT: The Noisy Channel Model

$$\text{argmax}_E \; p(E|F) = \text{argmax}_E \; p(F|E) \; p(E)$$

Why would we want to decompose $P(E|F)$ into $P(F|E)$ and $P(E)$?

- Model faithfulness and fluency explicitly
- $P(F|E)$ is learned from parallel corpus (for faithfulness)
- $P(E)$ can be learned from large monolingual corpus (for fluency)
- Similar techniques are used for other tasks as well (e.g., spell checkers, speech recognition)

Statistical MT and ASR

7. Machine Translation

Even though the problem of speech recognition remains unsolved to this day, some of us started to wonder in the mid 1980s whether our ASR methods could be successfully applied to new fields. Bob Mercer and I spent many of our after-lunch "periphery" walks discussing possible candidates. We soon came up with two: machine translation and stock market modeling. It is probably only coincidence that Bob eventually ended up investigating the possibilities of stock value prediction. Indeed, he and Peter Brown departed IBM in 1993 to work for the phenomenally successful hedge fund Renaissance Technologies. Eventually at least 10 former members of the IBM CSR group were to be employed by that same company. The performance of the Renaissance fund is legendary, but I have no idea whether any methods we pioneered at IBM have ever been used. My former colleagues will not tell me: Theirs is a very hush-hush operation!

On the other hand, we did start working on machine translation (MT) in 1987. As expected, we formulated the problem statistically. The basic diagram of MT, shown in Figure 4, is practically identical to that of ASR. In fact, even the basic formulas are identical except for a change in the letters that designate the variables:

$$\widehat{E} = \arg\max_{E} P(F|E)\, P(E)$$

Fred Jelinek. The Dawn of Statistical ASR and MT. 2009

A Translation Model needs to address:

Lexical translation problems

Reordering problems

- Local reordering
- Long-distance reordering

The Lexical Translation Model: IBM Model 1

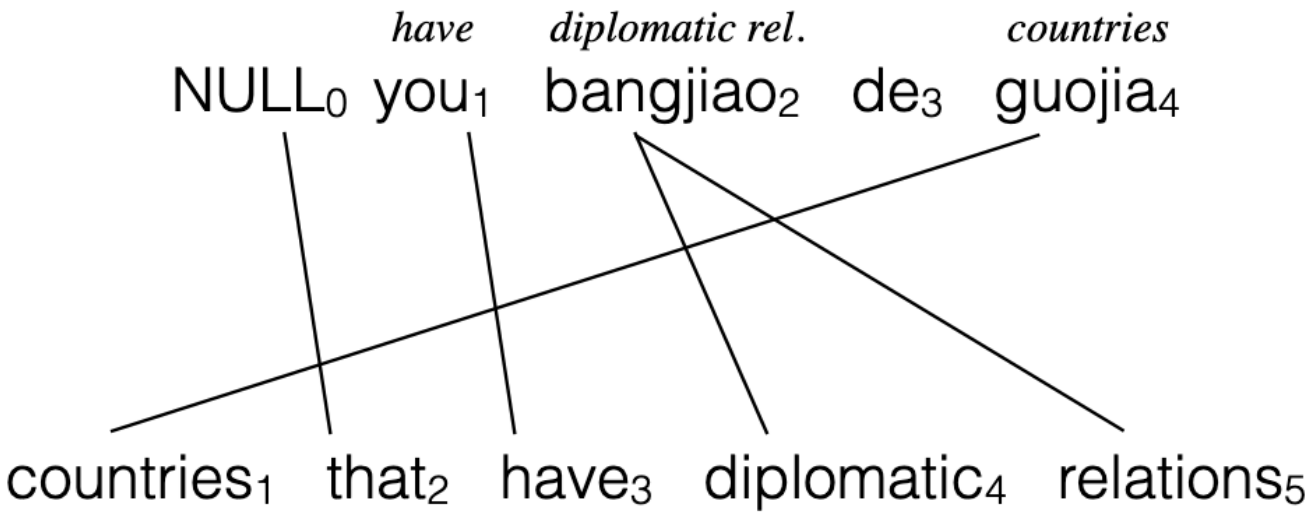
Translation is based on alignment

- Mapping a target word at position i to a source word at position j
  - alignment: $A : i \rightarrow j$

Models 1's one-many assumptions

- Each target word is generated by exactly one source word
- A target word can be generated by a NULL word; multiple target words can be generated by the same source word

Alignment in MT



$$A : \{1 \to 4, 2 \to 0, 3 \to 1, 4 \to 2, 5 \to 2\}$$

The Lexical Translation Model: IBM Model 1's Generative Story

Translation model generates a sentence of F (given E) in 3 steps:

- pick a length for F
- pick an alignment of F (length J) and E (length $I + 1$)

$$P(A|E) = \frac{\epsilon}{(I+1)^J}$$

- pick the j^{th} word of F based on English word e_i with which it aligns using distribution $t(f_j|e_i)$

$$P(F|E,A) = \prod_{j=1}^{J} t(f_j|e_{a_j})$$

The Lexical Translation Model: IBM Model 1's Generative Story

Combining previous equations, we have:

$$P(F, A|E) = P(F|E, A) \times P(A|E)$$

$$= \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} t(f_j|e_{a_j})$$

$$P(F|E) = \sum_A P(F, A|E) \qquad P(A|F, E) = \frac{P(F, A|E)}{\sum_A P(F, A|E)}$$

The heart of the translation model is the word translation probabilities $t(f_j|e_i)$

Finding the best alignment between a pair of sentences (i.e., alignment decoding) can be done efficiently in polynomial time

The Lexical Translation Model: EM training

We only have sentence-aligned (but not word-aligned) data

The EM procedure (we don't have the model and only have incomplete data)

- begins by assuming all word translations $t(f_j|e_i)$ are equally likely
- compute probabilities of alignments $P(A|F,E)$ given word translation probabilities (E-step 1)
- compute counts of aligned word pairs $tcount(f_j|e_i)$, weighted by alignment probabilities (E-step 2)
- recompute MLE word translation probabilities from these counts (M-step)
- repeat

Example

Considering a corpus with two sentences:

green house	the house
casa verde	la casa

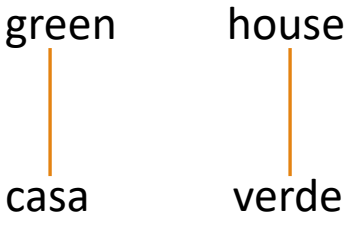
The vocabularies for the two languages are E = {green, house, the} and S = {casa, la, verde}

We start with uniform probabilities:

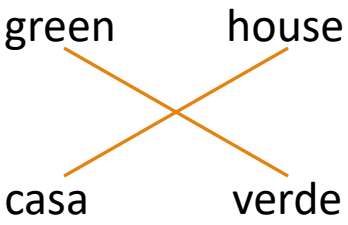
t(casa|green) = 1/3	t(verde|green) = 1/3	t(la|green) = 1/3
t(casa|house) = 1/3	t(verde|house) = 1/3	t(la|house) = 1/3
t(casa|the) = 1/3	t(verde|the) = 1/3	t(la|the) = 1/3

Example

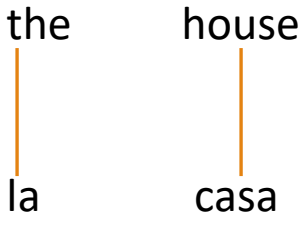
E-step 1a: We first compute $P(A, F|E)$ by multiplying all the t probabilities:



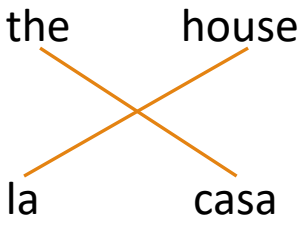
$P(A, F|E)$ =
t(casa|green)
x t(verde|house)
= 1/3x1/3
=1/9



$P(A, F|E)$ =
t(casa|house)
x t(verde|green)
= 1/3x1/3
=1/9



$P(A, F|E)$ =
t(la|the)
x t(casa|house)
= 1/3x1/3
=1/9

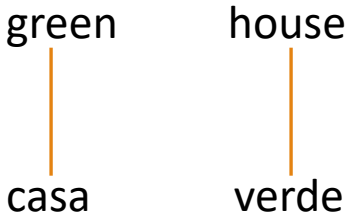


$P(A, F|E)$ =
t(la|house)
x t(casa|the)
= 1/3x1/3
=1/9

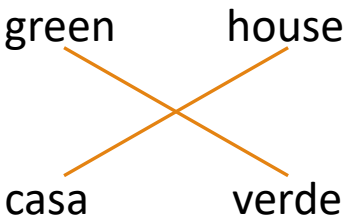
Example

E-step 1b: Normalize $P(A, F|E)$ to get $P(A|E, F)$ using:

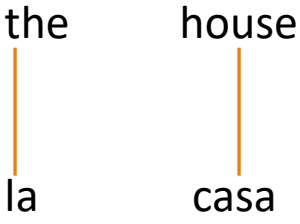
$$P(A|F, E) = \frac{P(F, A|E)}{\sum_A P(F, A|E)}$$



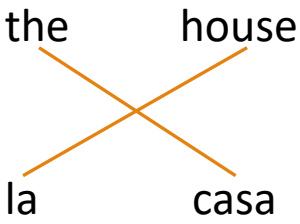
$$P(A|E, F) = \frac{1/9}{2/9} = \frac{1}{2}$$



$$P(A|E, F) = \frac{1/9}{2/9} = \frac{1}{2}$$



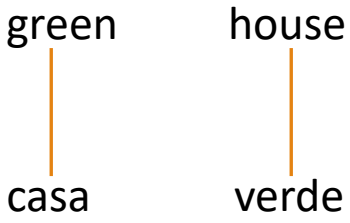
$$P(A|E, F) = \frac{1/9}{2/9} = \frac{1}{2}$$

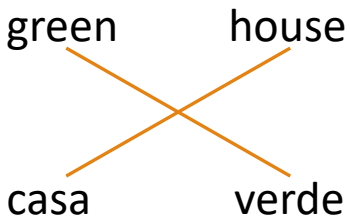


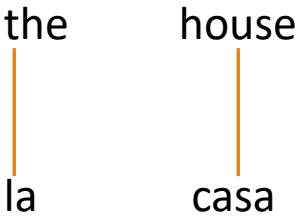
$$P(A|E, F) = \frac{1/9}{2/9} = \frac{1}{2}$$

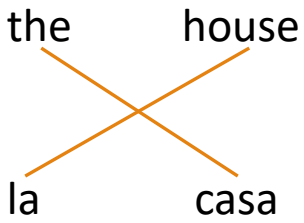
Example

E-step 2: Compute expected (fractional) counts of aligned word pairs, by weighting each count by $P(A|E,F)$:









$$P(A|E,F) = \frac{1/9}{2/9} = \frac{1}{2}$$

$$P(A|E,F) = \frac{1/9}{2/9} = \frac{1}{2}$$

$$P(A|E,F) = \frac{1/9}{2/9} = \frac{1}{2}$$

$$P(A|E,F) = \frac{1/9}{2/9} = \frac{1}{2}$$

tcount(casa|green) = 1/2	tcount(verde|green) = 1/2	tcount(la|green) = 0	total(green)=1
tcount(casa|house) = 1/2+1/2	tcount(verde|house) = 1/2	tcount(la|house) = 1/2	total(house)=2
tcount(casa|the) = 1/2	tcount(verde|the) = 0	tcount(la|the) = 1/2	total(the)=1

Example

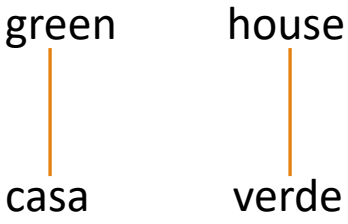
M-step: Compute the MLE probability parameters by normalizing the tcounts to sum to 1.

tcount(casa|green) = 1/2	tcount(verde|green) = 1/2	tcount(la|green) = 0	total(green)=1
tcount(casa|house) = 1/2+1/2	tcount(verde|house) = 1/2	tcount(la|house) = 1/2	total(house)=2
tcount(casa|the) = 1/2	tcount(verde|the) = 0	tcount(la|the) = 1/2	total(the)=1

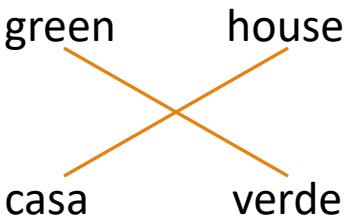
t(casa|green) = 1/2 / 1 = 1/2	t(verde|green) = 1/2 / 1 = 1/2	t(la|green) = 0/1 = 0
t(casa|house) = 1/2 = 1/2	t(verde|house) = 1/2 / 2 = 1/4	t(la|house) = 1/2 / 2 = 1/4
t(casa|the) = 1/2 / 1 = 1/2	t(verde|the) = 0/1 = 0	t(la|the) = 1/2 / 1 = 1/2

Example

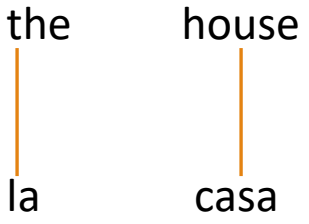
E-step 1a: We first compute $P(A, F|E)$ by multiplying all the t probabilities:



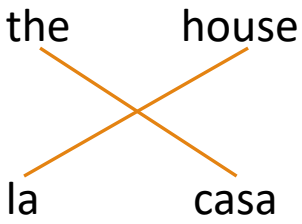
$P(A, F|E)$ =
t(casa|green)
x t(verde|house)
= 1/2x1/4
=1/8



$P(A, F|E)$ =
t(casa|house)
x t(verde|green)
= 1/2x1/2
=1/4



$P(A, F|E)$ =
t(la|the)
x t(casa|house)
= 1/2x1/2
=1/4



$P(A, F|E)$ =
t(la|house)
x t(casa|the)
= 1/2x1/4
=1/8

The Lexical Translation Model: EM training

We only have sentence-aligned (but not word-aligned) data

The EM procedure (we don't have the model and only have incomplete data)

- begins by assuming all word translations $t(fj|e_i)$ are equally likely
- compute probabilities of alignments $P(A|F,E)$ given word translation probabilities (E-step 1)
- compute counts of aligned word pairs $tcount(fj|e_i)$, weighted by alignment probabilities (E-step 2)
- recompute MLE word translation probabilities from these counts (M-step)
- repeat

Apply model to data

Estimate model from data

A Translation Model Needs To Address:

We just covered: Lexical translation problems

Reordering problems

- Local reordering
- Long-distance reordering

The Phrasal Translation Model

Word-based translation models are not good at capturing reordering

Phrase-based models tackles the local reordering problem by memorizing phrases

The “Standard Model” used by Google (before deep learning).

The Phrasal Translation Model

'Phrases' are word sequences that are consistent with the word alignment, i.e. not limited to linguistic phrases

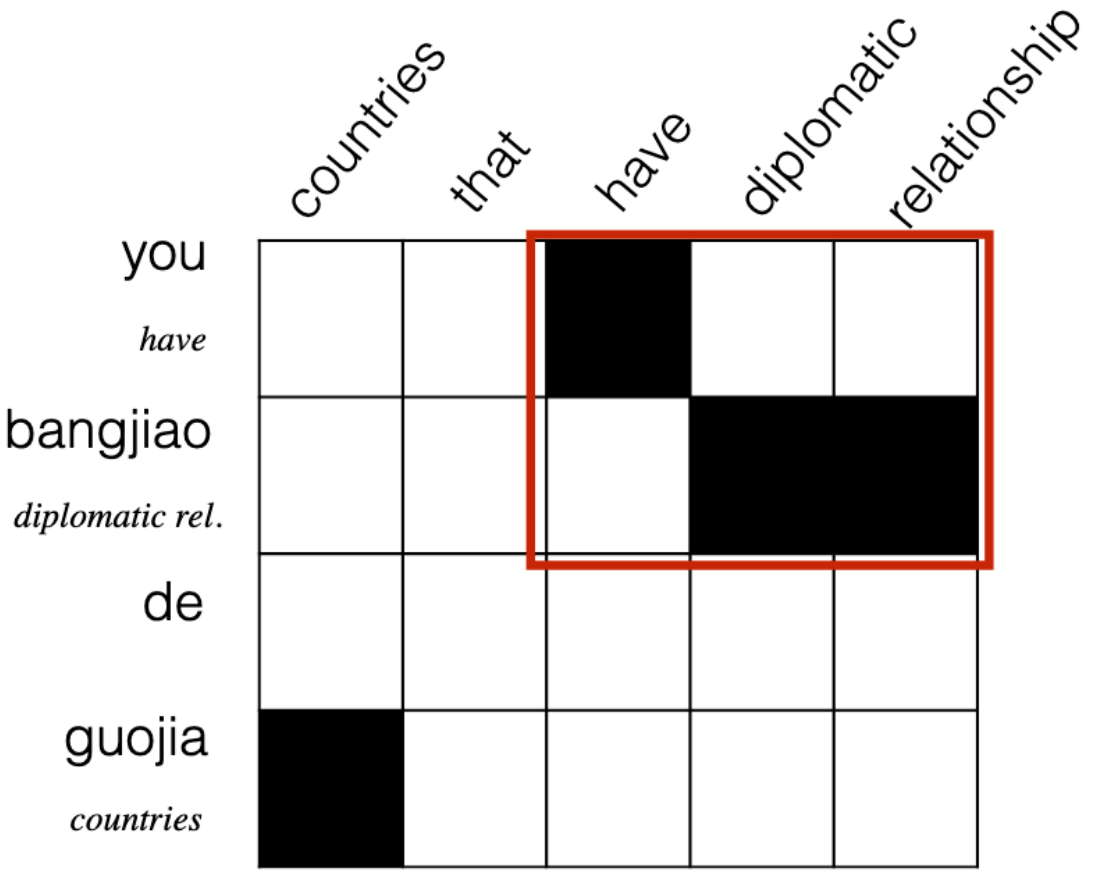
From parallel data, obtain word alignment A (using IBM Models etc.)

Extract all phrase pairs that are consistent with A

A phrase pair (E, F) is consistent with A, if each word e in E is either aligned to either a word f in F or NULL and vice versa, and all the words are only aligned with each other in the phrase pair and not to any external words.

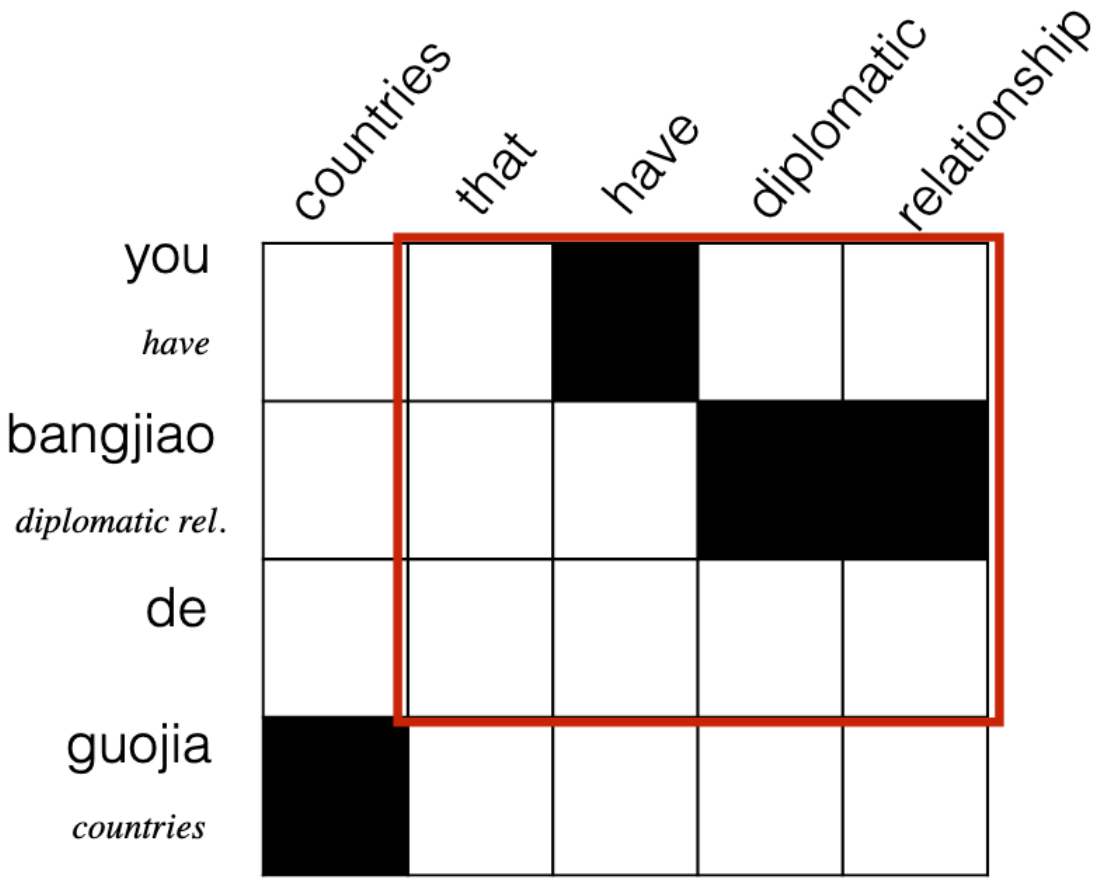
Score phrase pairs by relative frequency $\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$

The Phrasal Translation Model



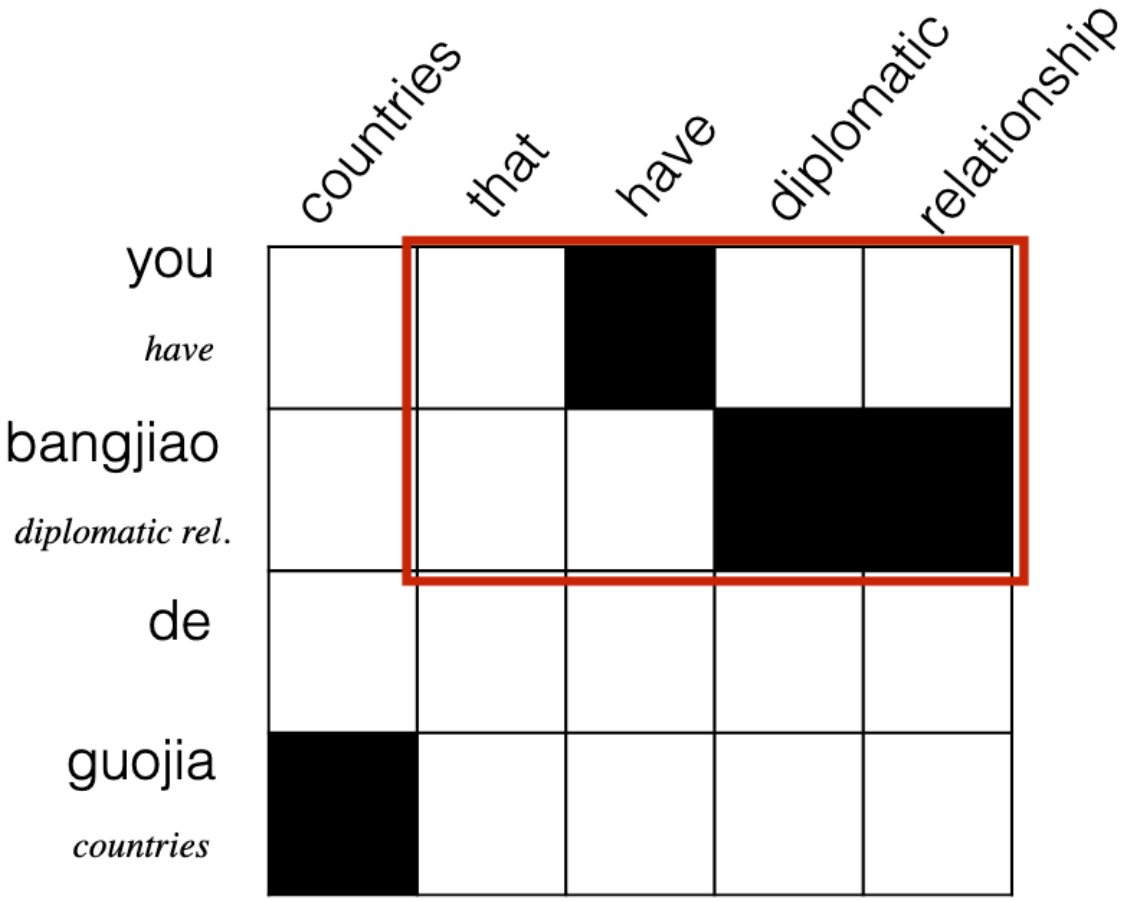
you bangjiao → have diplomatic relationship

The Phrasal Translation Model



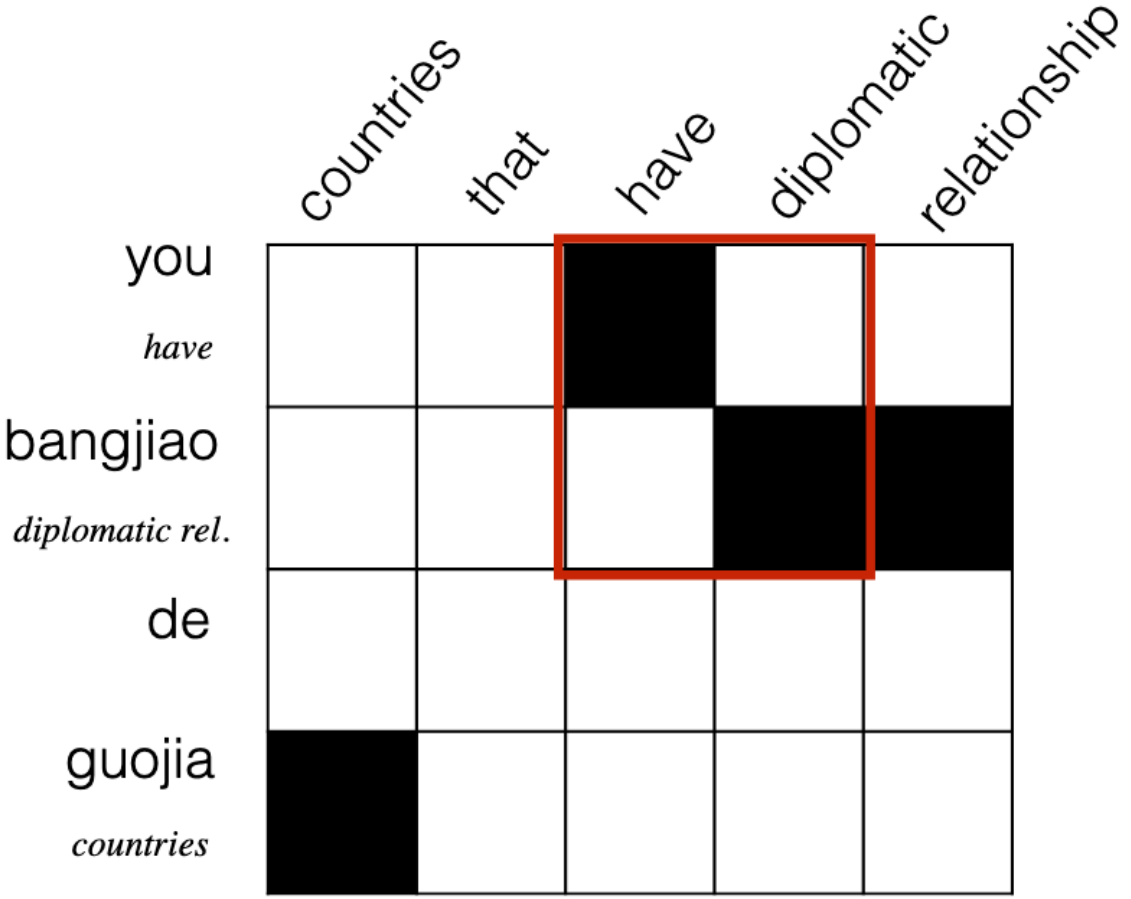
you bangjiao de → that have diplomatic relationship

The Phrasal Translation Model



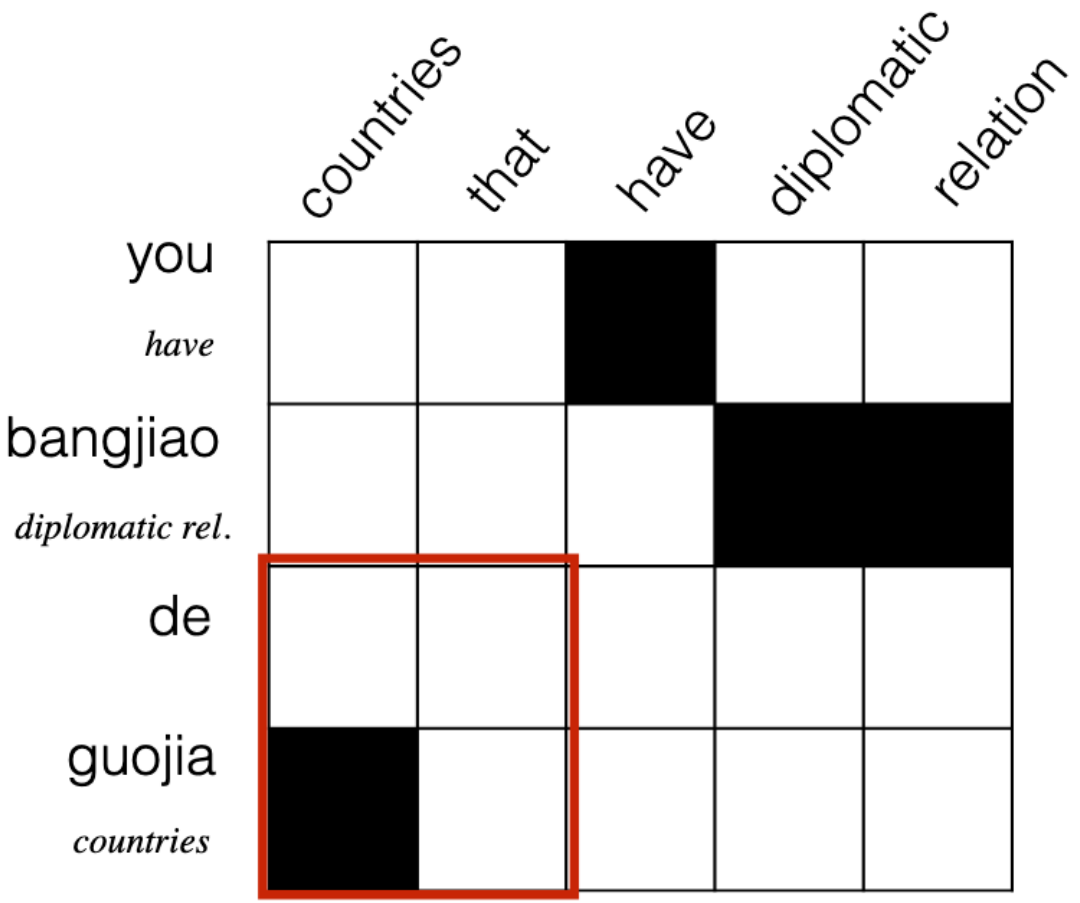
you bangjiao → that have diplomatic relationship

The Phrasal Translation Model



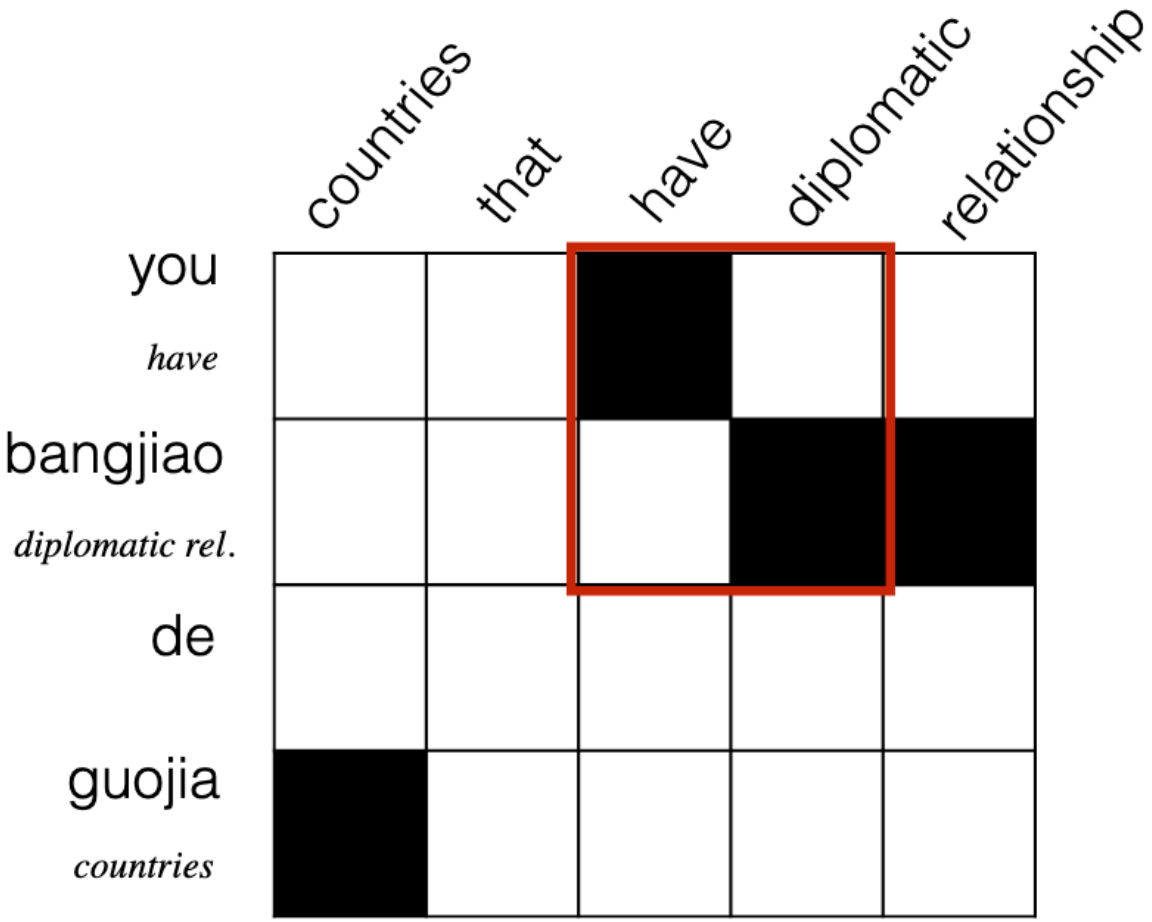
you bangjiao → have diplomatic

The Phrasal Translation Model



de guojia → countries that

The Phrasal Translation Model



Incorrect:

(*) you bangjiao → have diplomatic

A Translation Model needs to address:

Lexical translation problems

Reordering problems

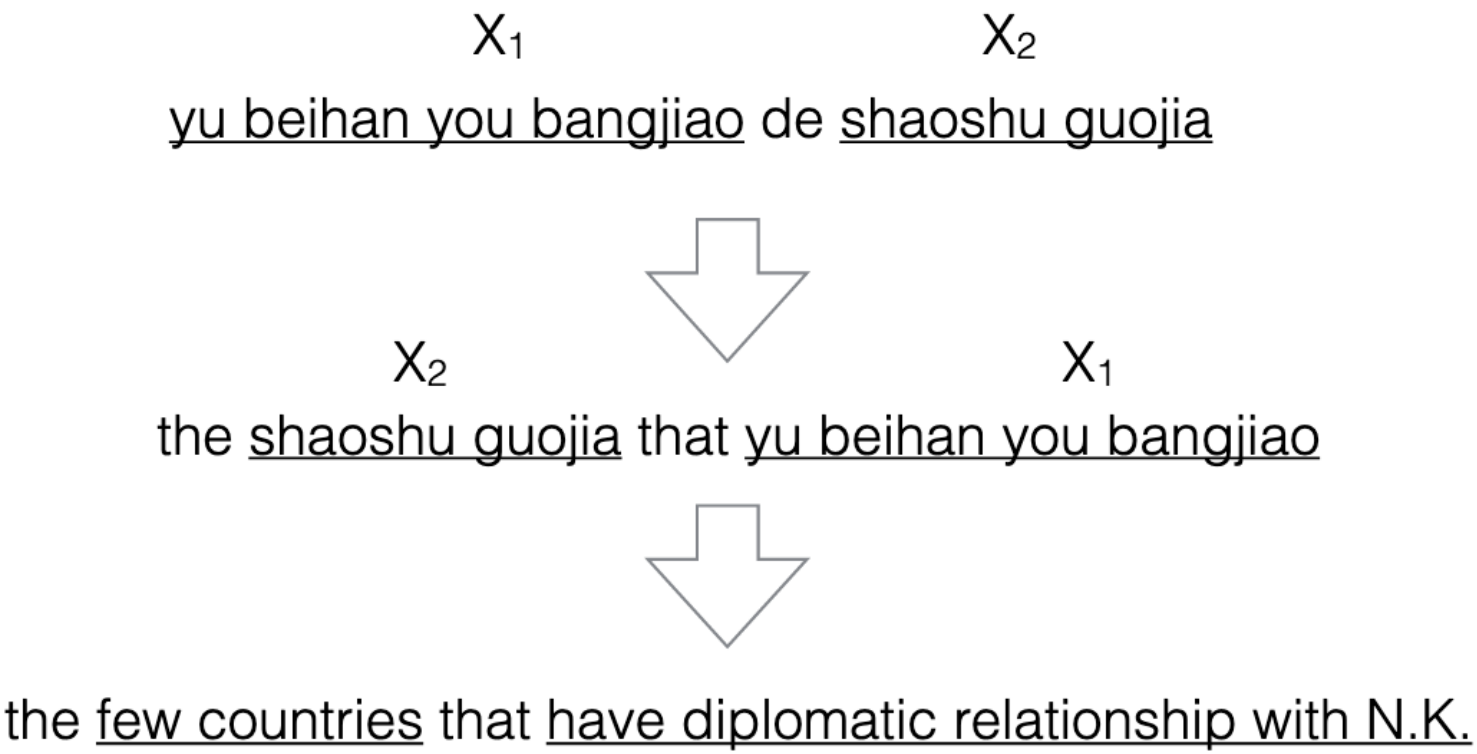
- **We just covered: Local reordering**
- Long-distance reordering

Log-linear and Syntax-informed Translation Models

Log-linear model: incorporating more features than $P(F|E)$ and $P(E)$

- e.g., distortion scores that reflect the quality of reordering (Och and Ney, 2002)
- the distortion probability that measures the probability of two consecutive in the source language phrases being separated in the target language by a span (of the target language words) of a particular length (e.g., $d(start_i -$

Syntax-informed Translation Models



A Translation Model needs to address:

Lexical translation problems

Reordering problems

- **Local reordering**
- **We just covered: Long-distance reordering**

The Language Modeling for translation fluency (i.e., $P(E)$) has been covered in the prior classes (e.g., with the n-gram models)!

Putting it Together: the Decoder

Once the translation model and language model have been trained, translation can be performed by evaluating

$$\text{argmax}_E = P(F|E)P(E)$$

Exhaustive search intractable; must perform pruning

- compute the probability of each partial translation
- extend only the b partial translations with the highest probability ('beam search')
- include in the probability of a partial translation (i.e., the current cost) a rough estimate of the probability associated with generating the rest of the sentence (i.e., the future cost)
- approximating the future cost by ignoring the distortion cost and just finding the sequence of foreign phrases that has the minimum product of the language model and translation model costs

Decoding

Stack decoding

Maintaining a stack of hypotheses.

- Actually a priority queue

Iteratively pop off the best-scoring hypothesis, expand it, put back on stack

The score for each hypothesis

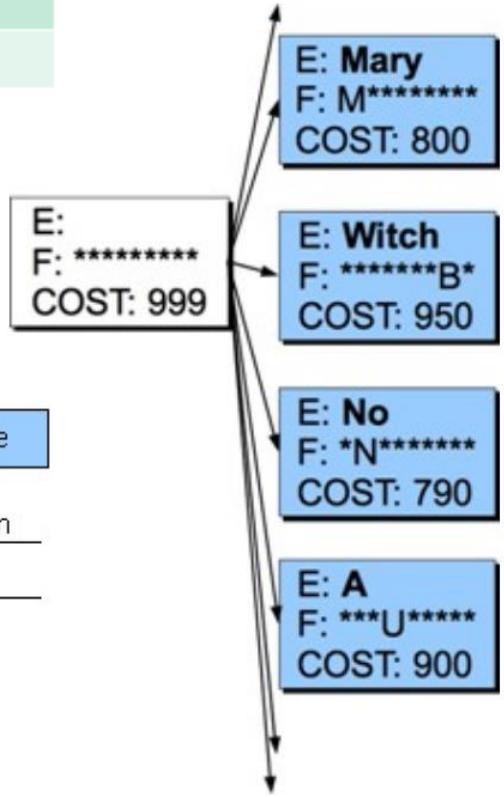
- The current score (the score so far):

$$cost(hyp(S(E,F)) = \prod_{i \in S} \phi(\overline{f}_i, \overline{e}_i) d(start_i - end_{i-1}) P(E)$$

- Estimate of future costs (approximation by Viterbi algorithm ignoring the distortion score)

Decoding: The Lattice Of Possible English Translations For Phrases

Position	1	2	3	4	5	6
English	Mary	did not	slap	the	green	witch
Spanish	Maria	no	dió una bofetada a	la	bruja	verde



Maria	no	dió	una	bofetada	a	la	bruja	verde

Mary	not	give	a	slap	to	the	witch	green
	did not		a slap		to		green witch	
	no	slap			to the			
	did not give				to			
					the			
		slap				the witch		

Evaluation of MT

Fluency: How intelligent, clear, readable, or natural in the target language is the translation?

Fidelity: Does the translation have the same meaning as the source?

- Adequacy: Does the translation convey the same information as source?
  - Bilingual judges given source and target language, assign a score
  - Monolingual judges given reference translation and MT result
- Informativeness: Does the translation convey enough information as the source to perform a task?
  - What % of questions can monolingual judges answer correctly about the source sentence given only the translation.

Automatic Evaluation of MT

Human evaluation is expensive and very slow

Need an evaluation metric that take seconds, not months

Intuition: MT is good if it looks like a human translation

1. Collect one or more human reference translations of the source.
2. Score MT output based on its similarity to the reference translations.

   - BLUE
   - NIST
   - TER
   - METEOR

BLUE (Bilingual Evaluation Understudy)

"n-gram precision"

Ratio of correct n-grams to the total number of output n-grams

- Correct: number of n-grams (unigram, bigram, etc.) the MT output shares with the reference translations.
- Total: number of n-grams in the MT result.

The higher precision, the better

Recall is ignored (as we don't know all the possible reference translations for a given sentence)

BLUE

Controlled by a number of parameters:

- N-gram order N. Most often $N = 4$
- Case sensitivity: By default, we compute case insensitive BLEU scores to evaluate a translator.
- Brevity, ρ, to penalize short translation.

Basically, the averaged percentage of n-gram matches.

Brevity penalty

$$BLEU_b = \rho \; BLEU_a \qquad \rho = \exp\{\min(0, \frac{n-L}{n})\} \qquad BLEU_a = \{\prod_{i=1}^{N} P(i)\}^{1/N}$$

$$P(i) = \frac{Matched(i)}{H(i)} \qquad Matched(i) = \sum_{t_i} \min\{C_h(t_i), \max_j C_{hj}(t_i)\}$$

Clipping to avoid rewarding candidates with extra repeated words

Where:

- t_i is an i-gram in the MT result
- $C_h(t_i)$ is the number of times t_i occurs in the result
- $C_{hj}(t_i)$ is the number of times t_i occurs in reference j
- $H(i)$ is the number of i-grams in the result ($H(i) = n - i + 1$)
- n is the length of the result
- L is the length of the reference (i.e., in case of multiple references, taking average, the shortest or the closet to n)

BLEU Example

Cand 1: Mary no slap the witch green
Cand 2: Mary did not give a smack to a green witch

Ref 1: Mary did not slap the green witch
Ref 2: Mary did not smack the green witch
Ref 3: Mary did not hit a green sorceress

Cand 1 Unigram Precision: 5/6

BLEU Example

Cand 1: Mary no slap the witch green
Cand 2: Mary did not give a smack to a green witch

Ref 1: Mary did not slap the green witch
Ref 2: Mary did not smack the green witch
Ref 3: Mary did not hit a green sorceress

Cand 1 Bigram Precision: 1/5

BLEU Example

Cand 1: Mary no slap the witch green
Cand 2: Mary did not give a smack to a green witch

Ref 1: Mary did not slap the green witch
Ref 2: Mary did not smack the green witch
Ref 3: Mary did not hit a green sorceress

Clip match count of each *n*-gram to maximum count of the *n*-gram in any single reference translation

Cand 2 Unigram Precision: 7/10

BLEU Example

Cand 1: Mary no slap the witch green
Cand 2: Mary did not give a smack to a green witch

Ref 1: Mary did not slap the green witch
Ref 2: Mary did not smack the green witch
Ref 3: Mary did not hit a green sorceress

Cand 2 Bigram Precision: 4/9

Modified *N*-Gram Precision

Average *n*-gram precision over all *n*-grams up to size *N* (typically 4) using geometric mean.

$$p = \sqrt[N]{\prod_{n=1}^{N} p_n}$$

Cand 1: $p = \sqrt[2]{\frac{5}{6}\frac{1}{5}} = 0.408$

Cand 2: $p = \sqrt[2]{\frac{7}{10}\frac{4}{9}} = 0.558$

BLEU Score

Final BLEU Score: BLEU = $BP \times p$

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$$

$$BLEU = 0.846 \times 0.408 = 0.345$$

Cand 2: Mary did not give a smack to a green witch.

Best Ref: Mary did not smack the green witch.

$$c = 10, \quad r = 7, \quad BP = 1$$

$$BLEU = 1 \times 0.558 = 0.558$$

BLEU Score Issues

BLEU has been shown to correlate with human evaluation when comparing outputs from different SMT systems.

However, it is does not correlate with human judgments when comparing SMT systems with manually developed MT (Systran) or MT with human translations.

Other MT evaluation metrics have been proposed that claim to overcome some of the limitations of BLEU.