Machine Translation

Dan Klein
UC Berkeley

Slides from John DeNero

Translation Task

- Text is both the input and the output.
- Input and output have roughly the same information content.
- Output is more predictable than a language modeling task.
- Lots of naturally occurring examples (but not much metadata).

Translation Examples

English-German News Test 2013 (a standard dev set)

```
Republican leaders justified their policy by the need
to combat electoral fraud.

  Die   Führungskräfte  der     Republikaner
   |          |          |            |
  The   Executives      of the  republican

  rechtfertigen  ihre  Politik   mit   der
        |         |       |       |     |
     justify     your  politics  With  of the

  Notwendigkeit  ,  den  Wahlbetrug      zu
        |        |   |       |           |
       need      ,  the  election fraud  to

  bekämpfen  .
      |      |
    fight    .
```

Variety in Human-Generated Translations

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomists got to know this incident 4 days later. This small planet is 50m in diameter. The astonomists are hard to find it for it comes from the direction of sun.

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

From https://catalog.ldc.upenn.edu/LDC2003T17

An asteroid that was large enough to destroy a medium-

Variety in Machine Translations

Human-generated reference translation

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomists got to know this incident 4 days later. This small planet is 50m in diameter. The astonomists are hard to find it for it comes from the direction of sun.

A commercial system from 2002

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

Google Translate, 2020

An asteroid that was large enough to destroy a medium-

From https://catalog.ldc.upenn.edu/LDC2003T17

Evaluation

BLEU Score

BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty (harshly penalizes translations shorter than the reference).

$$\text{Matched}_i = \sum_{t_i} \min\left\{C_h(t_i), \max_j C_j(t_i)\right\}$$

If "of the" appears twice in hypothesis h but only at most once in a reference, then only the first is "correct"

$$P_i = \frac{\text{Matched}_i}{H_i}$$

"Clipped" precision of n-gram tokens

$$B = \exp\left\{\min\left(0, \frac{n-L}{n}\right)\right\}$$

Brevity penalty only matters if the hypothesis **corpus** is shorter than the shortest reference.

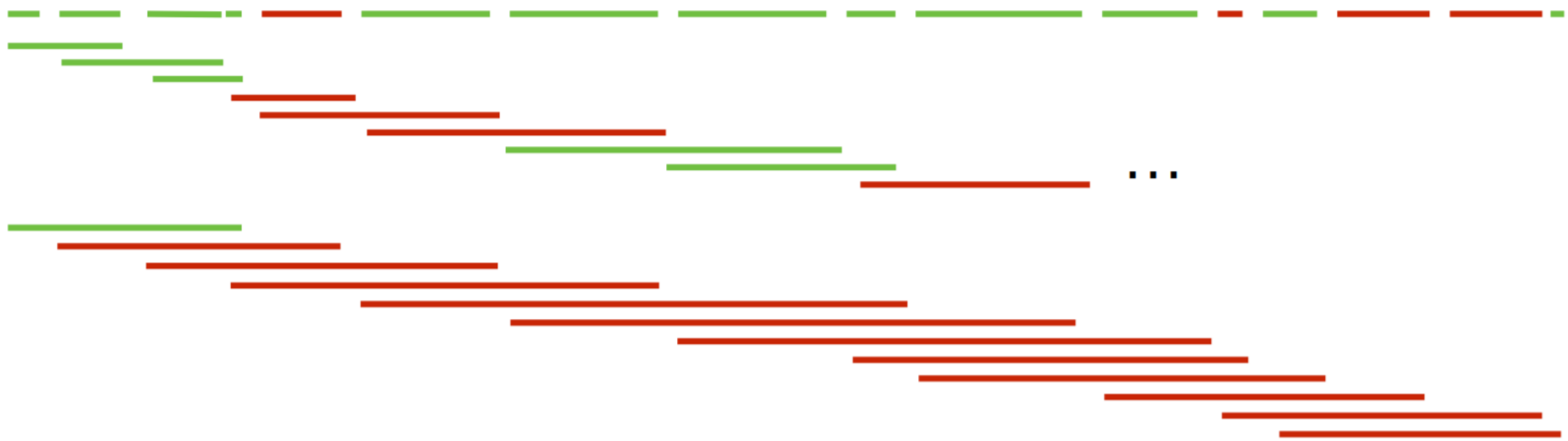
$$\text{BLUE} = B\left(\prod_{i=1}^{4} P_i\right)^{\frac{1}{4}}$$

BLEU is a mean of clipped precisions, scaled down by the brevity penalty.

Evaluation with BLEU

In this sense, the measures will partially undermine the American democratic system.

In this sense, these measures partially undermine the democratic system of the United States.



BLEU = 26.52, 75.0/40.0/21.4/7.7 (BP=1.000, ratio=1.143, hyp_len=16, ref_len=14)

(Papineni et al., 2002) BLEU: a method for automatic evaluation of machine translation.

Corpus BLEU Correlations with Average Human Judgments

These are ecological correlations over multiple segments; segment-level BLEU scores are noisy.

Commercial machine translation providers seem to all perform human evaluations of some sort.

(Ma et al., 2019) Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges

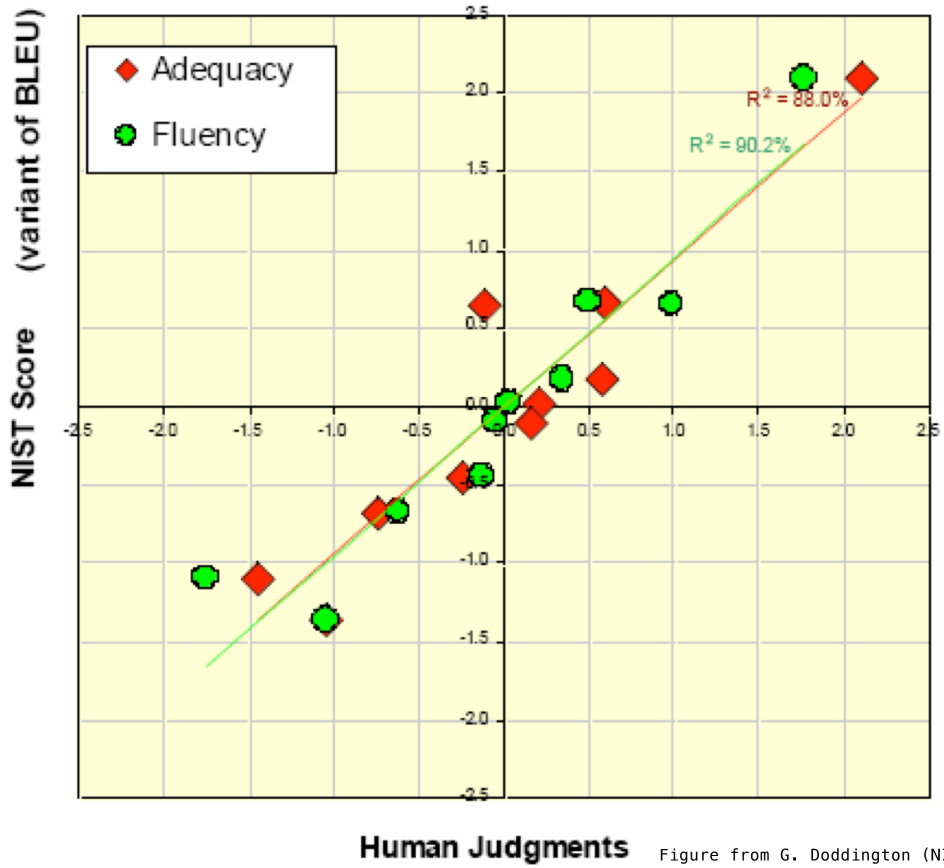


Figure from G. Doddington (NIST)

Human Evaluations

Direct assessment: adequacy & fluency

- Monolingual: Ask humans to compare machine translation to a human-generated reference. (Easier to source annotators)
- Bilingual: Ask humans to compare machine translation to the source sentence that was translated. (Compares to human quality)
- Annotators can assess segments (sentences) or whole documents.
- Segments can be assessed with or without document context.

Ranking assessment:

- Raters are presented with 2 or more translations.
- A human-generated reference may be provided, along with the source.
- "In a pairwise ranking experiment, human raters assessing adequacy and fluency show a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences." (Laubli et al., 2018)

Editing assessment: How many edits required to reach human quality

(Laubli et al., 2018) Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

Translationese and Evaluation

Translated text can: (Baker et al., 1993; Graham et al., 2019)

- be more explicit than the original source
- be less ambiguous
- be simplified (lexical, syntactically and stylistically)
- display a preference for conventional grammaticality
- avoid repetition
- exaggerate target language features
- display features of the source language

"If we consider only original source text (i.e. not translated from another language, or translationese), then we find evidence showing that human parity has not been achieved."
(Toral et al., 2018)

(Baker et al., 1993) Corpus linguistics and transla- tion studies: Implications and applications.
(Graham et al., 2019) Translationese in Machine Translation Evaluation.
(Toral et al, 2018) Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation

WMT 2019 Evaluation

2019 segment-in-context direct assessment (Barrault et al, 2019):

- ✓ German to English: many systems are tied with human performance;
- × English to Chinese: all systems are outperformed by the human translator;
- × English to Czech: all systems are outperformed by the human translator;
- × English to Finnish: all systems are outperformed by the human translator;
- ✓ English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance;
- × English to Gujarati: all systems are outperformed by the human translator;
- × English to Kazakh: all systems are outperformed by the human translator;
- × English to Lithuanian: all systems are outperformed by the human translator;
- ✓ English to Russian: Facebook-FAIR is tied with human performance.

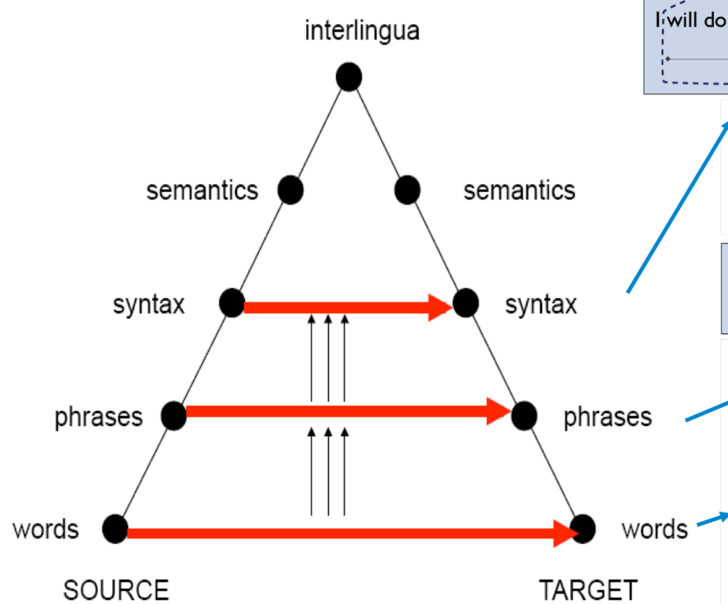
(Barrault et al, 2019) Findings of the 2019 Conference on Machine Translation (WMT19)

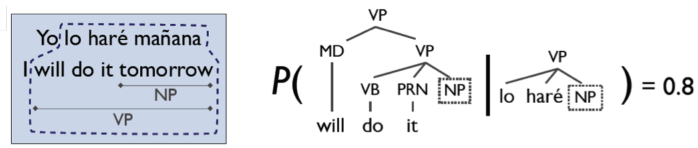
Statistical Machine Translation
(1990 - 2015)

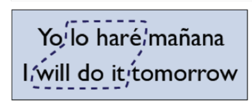
When I look at an article in Russian, I say: “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

Warren Weaver (1949)

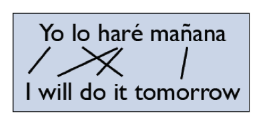
Levels of Transfer: Vauquois Triangle (1968)





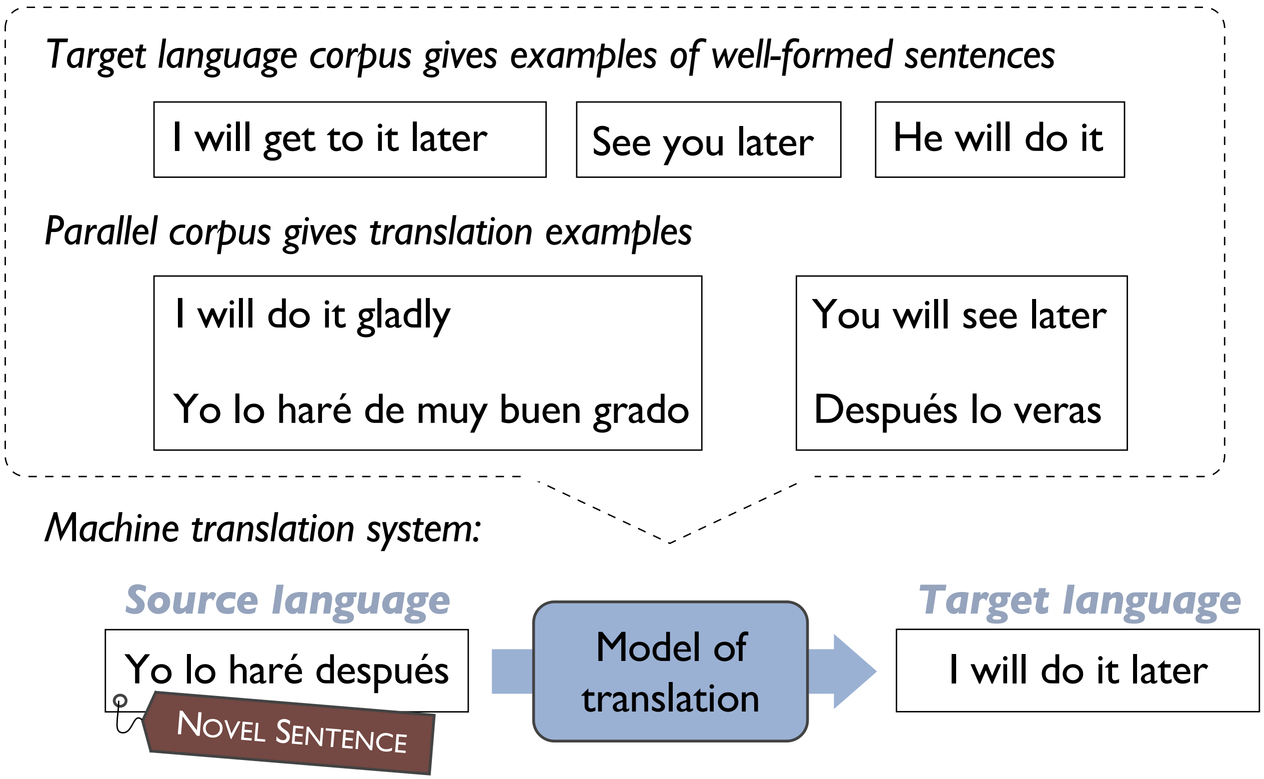


English (E)	*P*(E | lo haré)
will do it	0.8
will do so	0.2

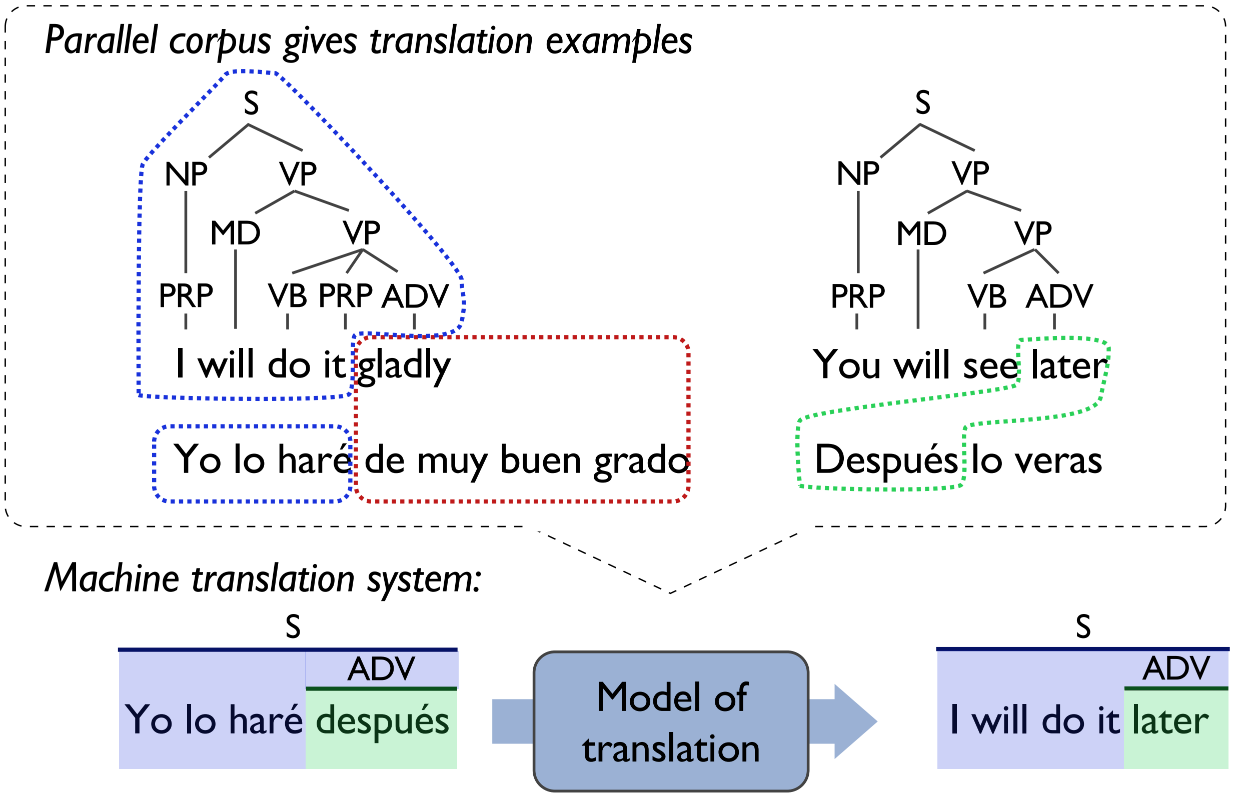


English (E)	*P*(E | mañana)
tomorrow	0.7
morning	0.3

Data-Driven Machine Translation



Stitching Together Fragments



Evolution of the Noisy Channel Model

$$P(e|f) \propto P(f|e) \cdot P(e)$$

$$\max_e P(e|f) = \max_e P(f|e) \cdot P(e)$$

$$P(e|f) \propto P(f|e)^{\phi_{\mathrm{tm}}} \cdot P(e)^{\phi_{\mathrm{lm}}}$$

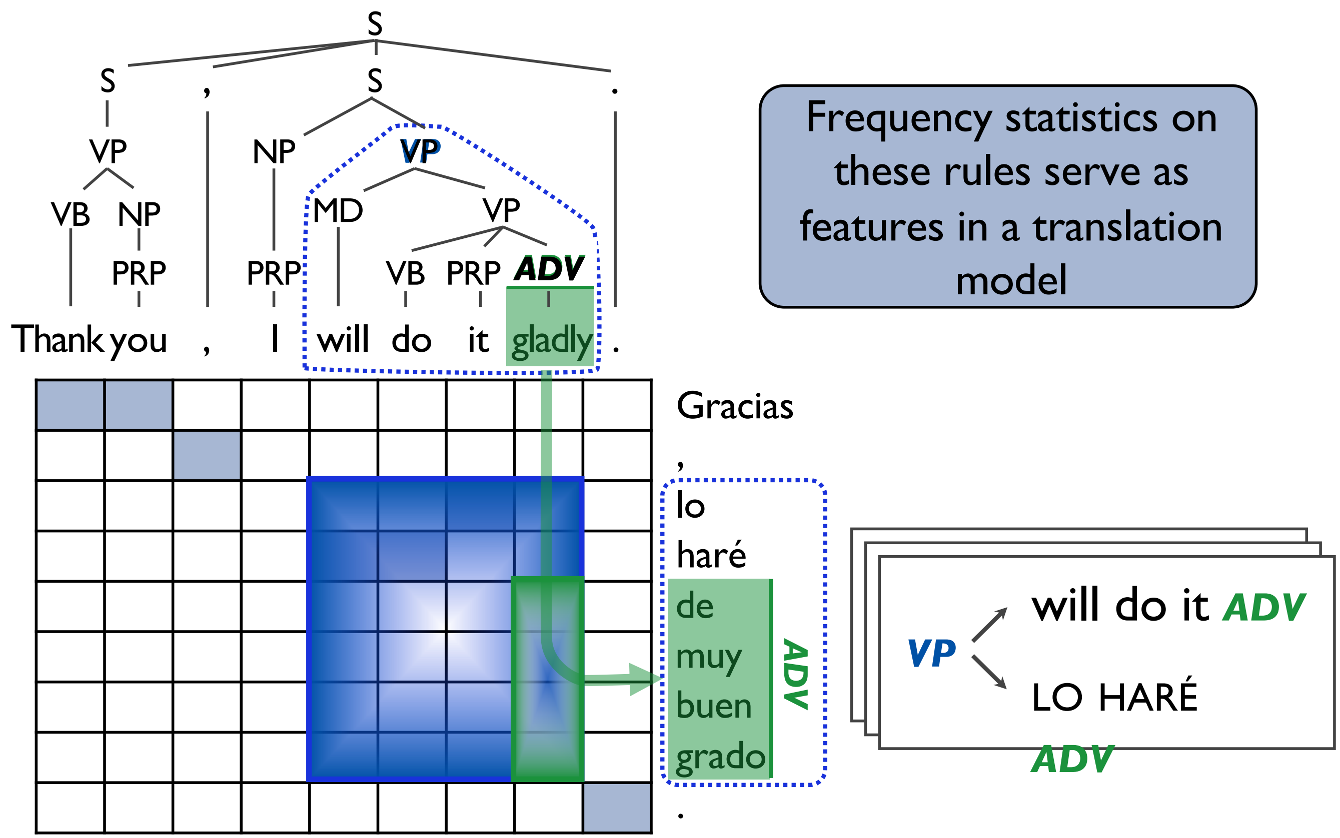
Chosen to minimize loss

$$P(e|f) \propto \exp\left\{\sum_i w_i \cdot f_i(e,f)\right\}$$

E.g., log P(e)

Word Alignment

Extracting Translation Rules



Frequency statistics on these rules serve as features in a translation model

Counting Aligned Phrases

d'assister à la reunion et ||| to attend the meeting and

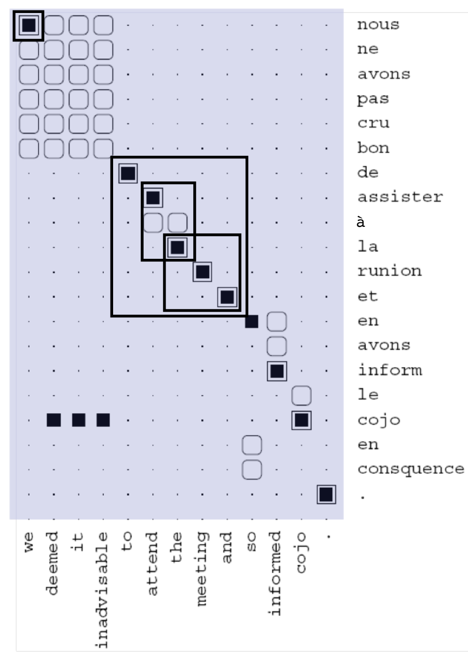
assister à la reunion ||| attend the meeting

la reunion and ||| the meeting and

nous ||| we

...

- Relative frequencies are the most important features in a phrase-based or syntax-based model.
- Scoring a phrase under a lexical model is the second most important feature.
- Estimation does not involve choosing among segmentations of a sentence into phrases.



Slide by Greg Durrett

Interlude: Lexical Translation Models

HMM Alignment Model

Alignment Link Posteriors

$$c(e|f;\mathbf{e},\mathbf{f}) = \sum_a \boxed{p(a|\mathbf{e},\mathbf{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})}$$

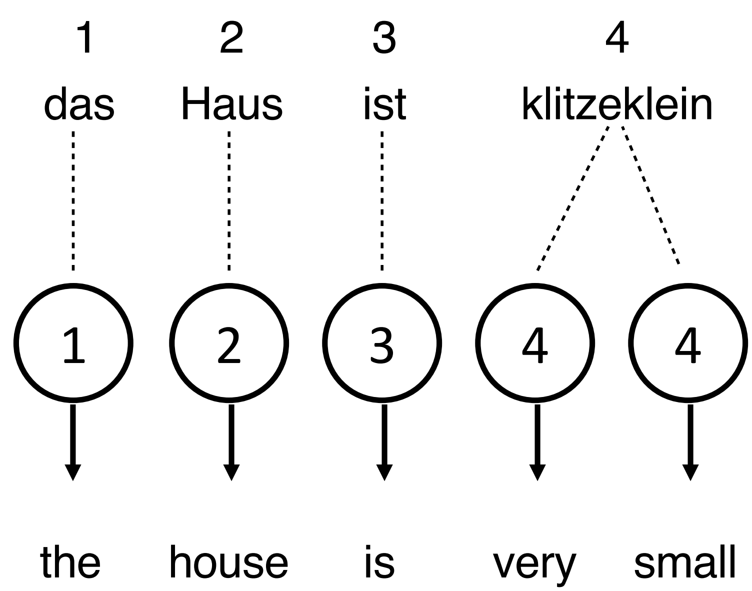
Non-zero for any alignment vector
(for sentence pair **e, f**)
that has word e aligned to word f

$$\begin{aligned} c(e|f;\mathbf{e},\mathbf{f}) &= \sum_i \sum_j \delta(e,e_j) \cdot \delta(f,f_i) \cdot P(a(j)=i|\mathbf{e},\mathbf{f}) \\ &= \sum_i \sum_j \delta(e,e_j) \cdot \delta(f,f_i) \cdot \sum_a \boxed{P(a|\mathbf{e},\mathbf{f}) \cdot \delta(a(j),i)} \end{aligned}$$

Non-zero for any alignment vector
(for sentence pair **e, f**)
that has position j aligned to position i

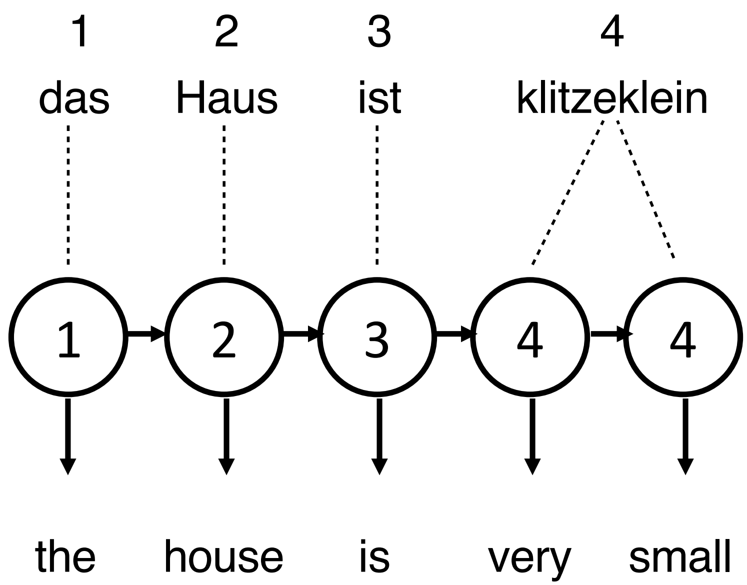
Model 1 Posteriors

$$P(a(j) = i|\mathbf{e}, \mathbf{f}) = \frac{t(e_j|f_i)}{\sum_{i'} t(e_j|f_{i'})}$$



HMM Alignment Model

$$P(a, \mathbf{e}|\mathbf{f}) \propto \prod_j P(e_j|f_{a(j)}) \cdot P(a(j)|a(j-1))$$



(Vogel, Stephan, Hermann Ney, and Christoph Tillmann, 1996) "HMM-based word alignment in statistical translation."
(Liang, Percy, Ben Taskar, and Dan Klein, 2006) "Alignment by agreement."

HMM Alignment Model Posteriors

$$P(a(j) = i|\mathbf{e}, \mathbf{f}) = \sum_a P(a|\mathbf{e}, \mathbf{f}) \cdot \delta(a(j), i)$$

Non-zero for alignments where j is aligned to i

$$= \sum_a \frac{P(a, \mathbf{e}|\mathbf{f}) \cdot \delta(a(j), i)}{P(\mathbf{e}|\mathbf{f})}$$

Words up to i (summing over alignments)

$$= \frac{\alpha_j(i) \cdot \beta_j(i)}{P(\mathbf{e}|\mathbf{f})}$$

Forward-Backward algorithm

$$\alpha_j(i) = \sum_{i'} P(a(j) = i|a(j-1) = i') \cdot P(e_j|f_i) \cdot \alpha_{j-1}(i')$$

Words after i

$$\beta_j(i) = \sum_{i''} P(a(j+1) = i''|a(j) = i) \cdot P(e_{j+1}|f_{i''}) \cdot \beta_{j+1}(i'')$$

$$\alpha_j(i) = P(e_1, e_2, \ldots, e_j, a(j) = i|\mathbf{f})$$

$$\beta_j(i) = P(e_{j+1}, e_{j+2}, \ldots, e_\ell|a(j) = i, \mathbf{f})$$

Interlude: Phrase-Based Models

What's Next?

Neural models: attention and the transformer architecture

Tricks of the trade: back-translation, knowledge distillation, subword models, and coverage vectors