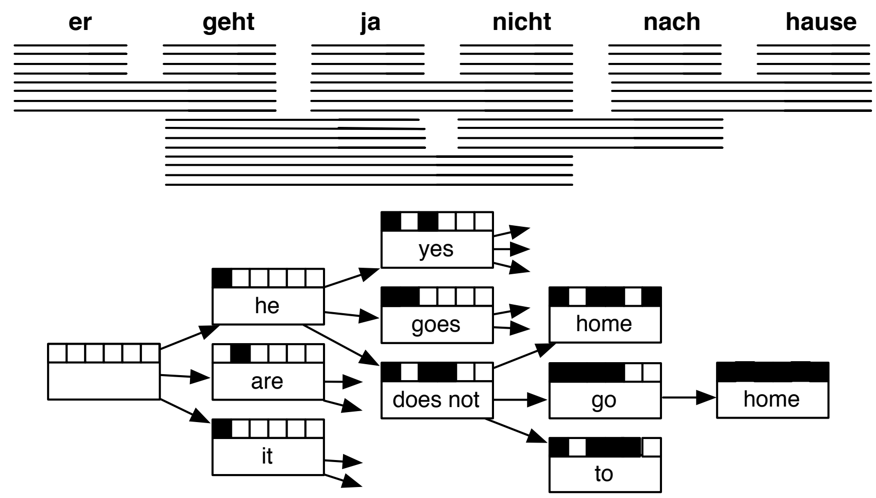


COMP90042 LECTURE 21

MT: PHRASE BASED & NEURAL ENCODER-DECODER

OVERVIEW

- Phrase based SMT
  - Scoring formula
  - Decoding algorithm
- Neural network ‘encoder-decoder’

WORD- AND PHRASE-BASED MT

- Seen word based models of translation
  - now used for *alignment*, but not actual *translation*
  - overly simplistic formulation
- Phrase based MT
  - treats n-grams as translation units, referred to as ‘phrases’ (not linguistic phrases though)

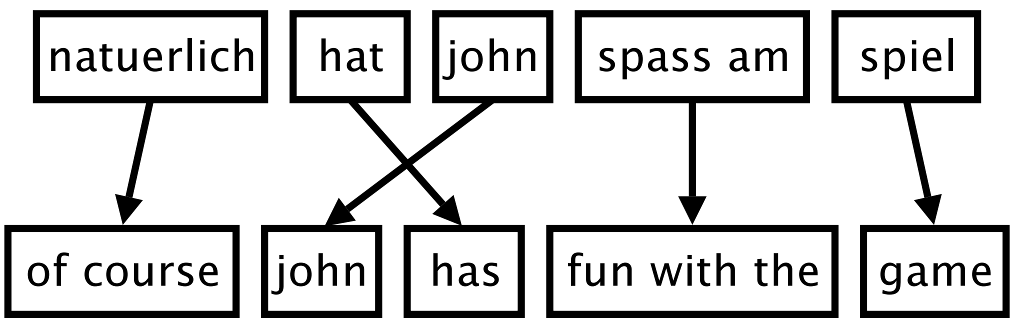
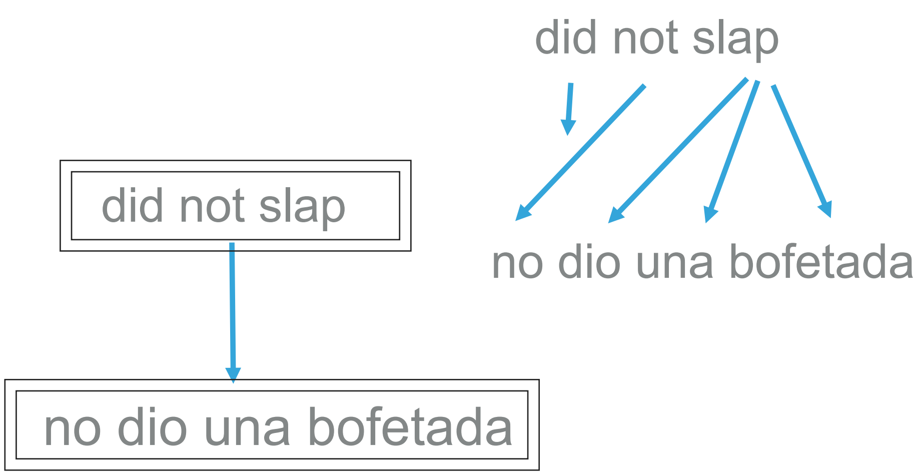
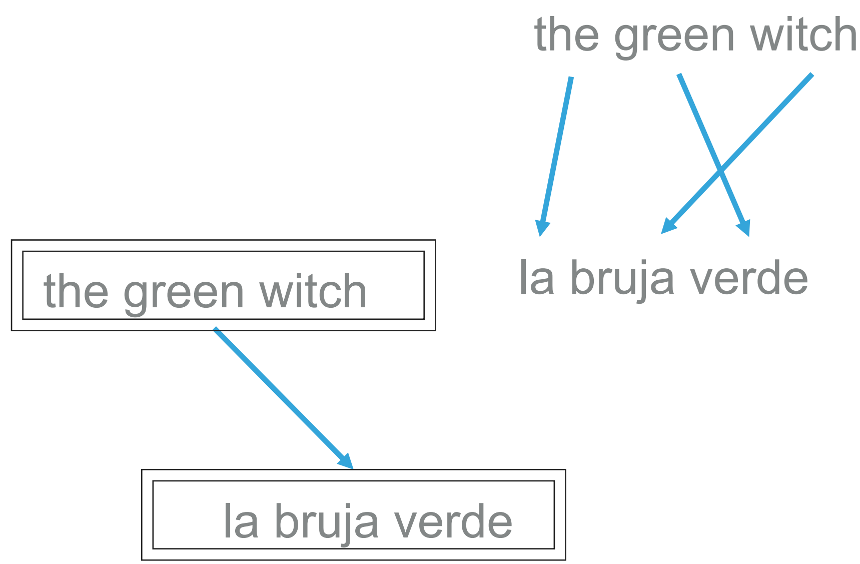


Fig from Koehn09

PHRASE VS WORD BASED MT

- Phrase-pairs memorise:
  - common translation fragments (have access to local context in choosing lexical translation)
  - common reordering patterns (making up for naïve models of reordering)





FINDING & SCORING PHRASE PAIRS

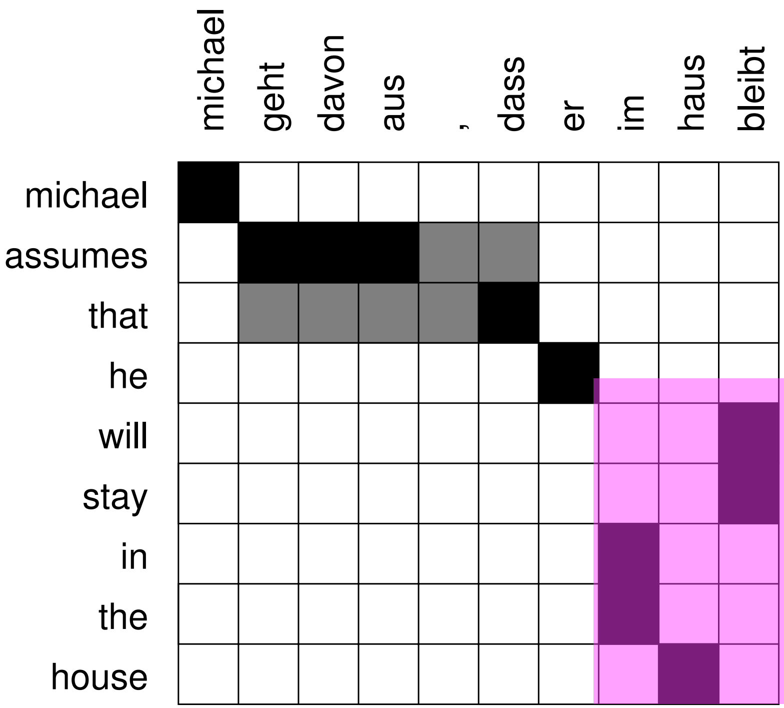


Fig from Koehn09

- “Extract” phrase pairs as contiguous chunks in word aligned text; then
  - compute counts over the whole corpus
  - normalise counts to produce ‘probabilities’
- E.g.,

$$\phi(\text{im haus bleibt}|\text{will stay in the house}) = \frac{c(\text{will stay in the house; im haus bleibt})}{c(\text{im haus bleibt})}$$

THE PHRASE-TABLE

- The **phrase-table** consists of all phrase-pairs and their scores, which forms the search space for decoding
  - E.g., for ***natuerlich*** it may contain the following translation phrases

Translation	**Probability** $p(e\|f)$
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

  - generally a massive list with many millions of phrase-pairs

DECODING

$$E^*, A^* = \text{argmax}_{E,A}\, \text{score}(E, A, F)$$

- *A* describes the segmentation of *F* into phrases; and the re-ordering of their translations to produce *E*

- The *score* function is a product of the
  - translation "probability", *P(F|E)*, split into phrase-pairs
  - language model probability, *P(E),* over full sentence E
  - distortion cost, $d(start_i, end_{i-1})$, measuring amount of reordering between adjacent phrase-pairs
- Search problem
  - find translation E* with the best overall score

TRANSLATION PROCESS

- Score the translations based on translation probabilities (step 2), reordering (step 3) and language model scores (steps 2 & 3).

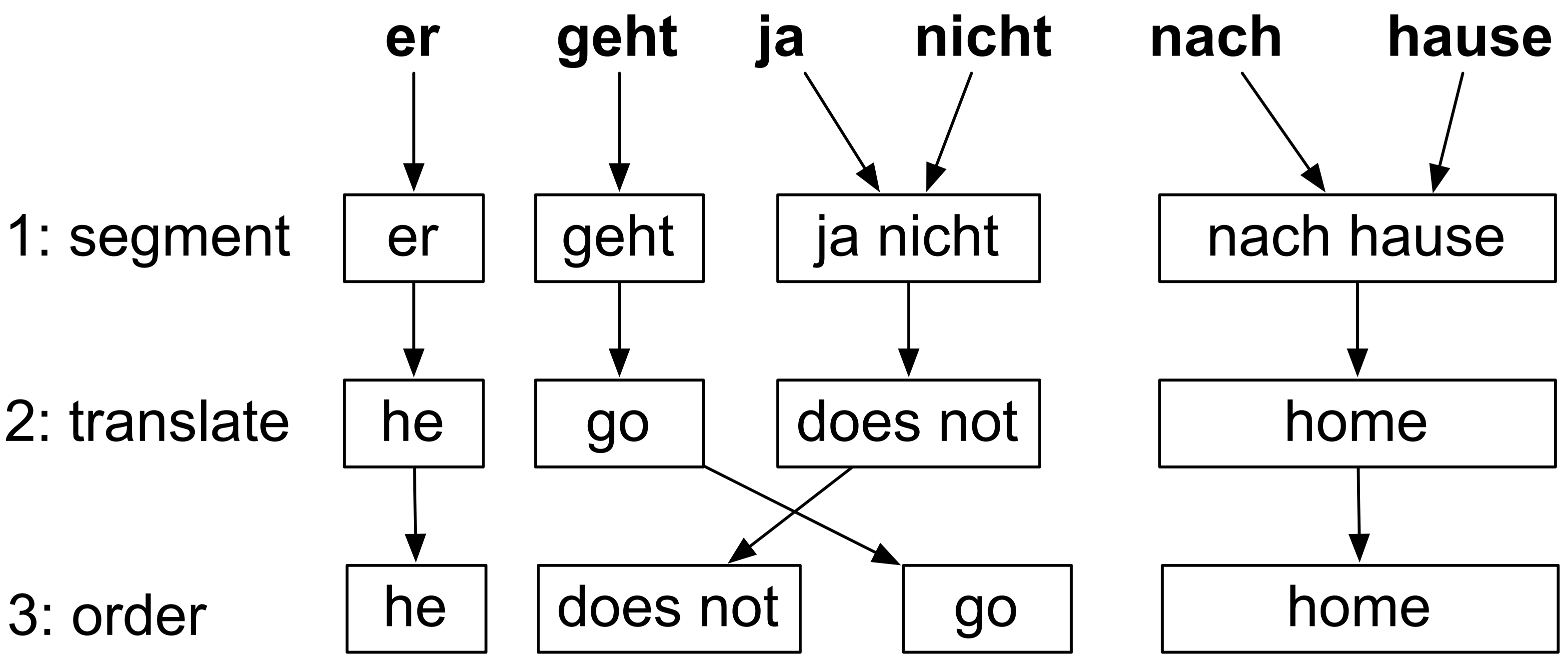
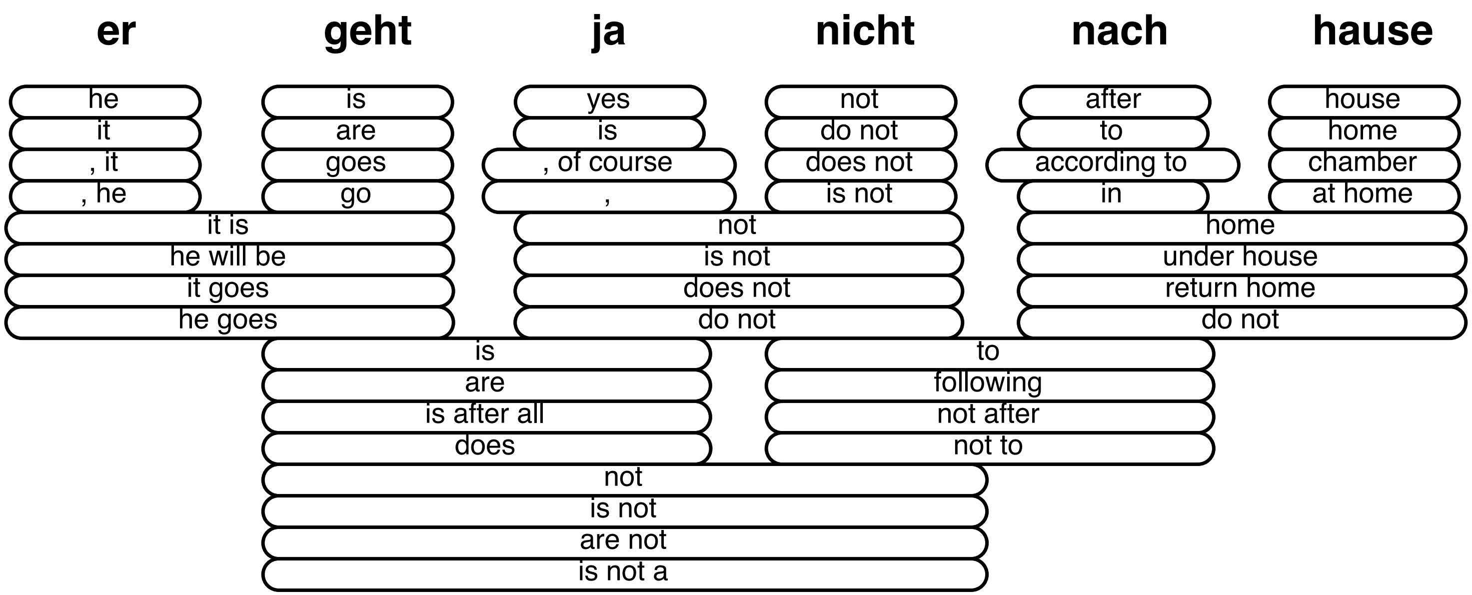


Figure from Koehn, 2009

SEARCH PROBLEM



- Cover all source words exactly once; visited in any order; and with any segmentation into "phrases"
- Choose a translation from phrase-table options

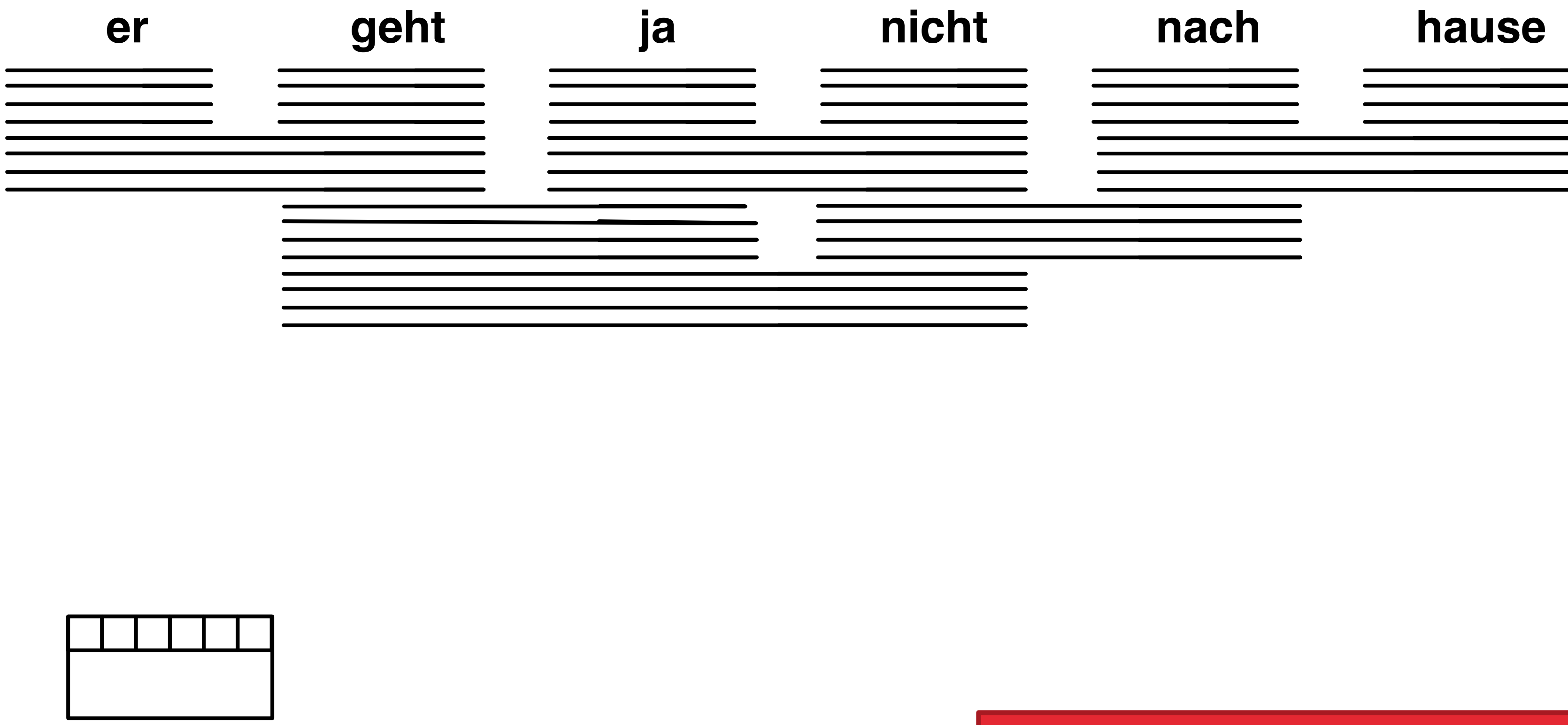
Leads to millions of possible translations…

Figure from Koehn, 2009

DYNAMIC PROGRAMMING SOLUTION

- Akin to Viterbi algorithm
  - factor out repeated computation
    (like Viterbi for HMMs, “chart” used in parsing)
  - efficiently solve the maximisation problem
- Aim is to translate every word of the input once
  - searching over *every* segmentation into phrases;
  - the translations of each phrase; and
  - all possible ordering of the phrases

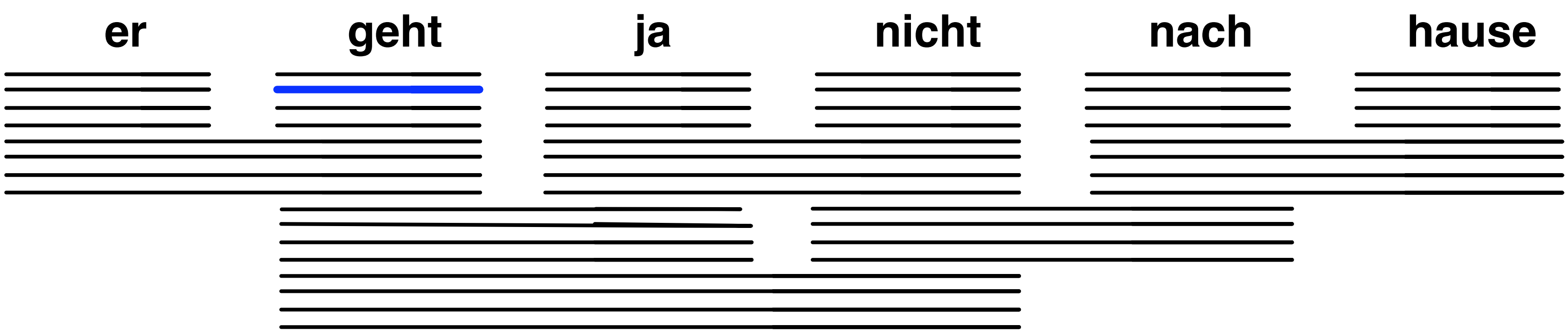
PHRASE-BASED DECODING

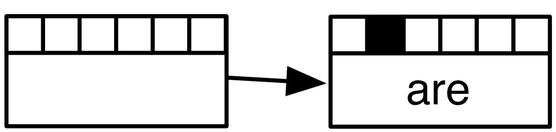


Start with empty state

Figure from Koehn, 2009

PHRASE-BASED DECODING

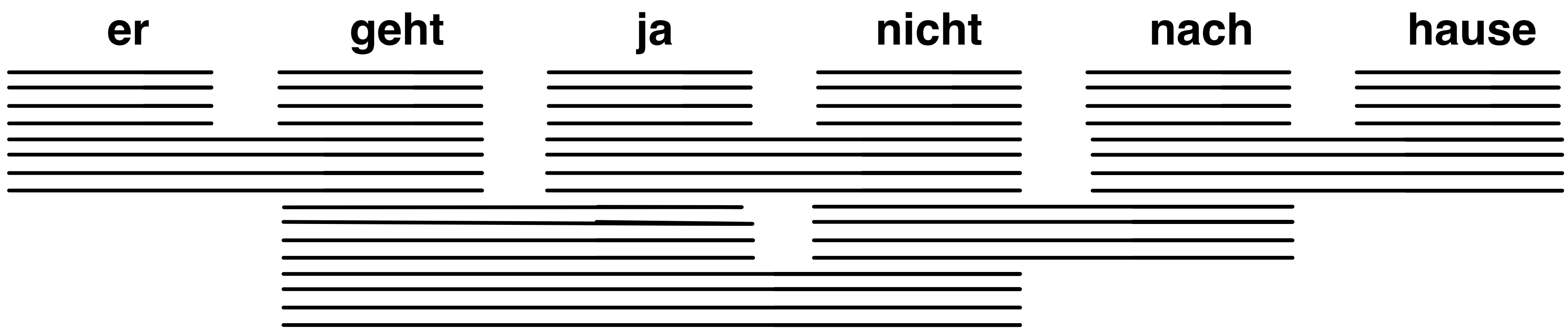


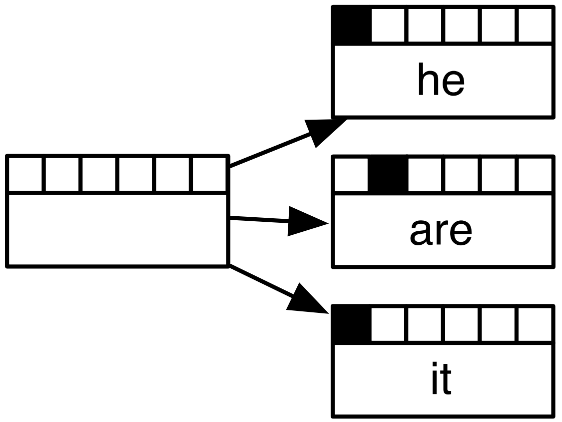


Expand by choosing input span and generating translation

Figure from Koehn, 2009

PHRASE-BASED DECODING

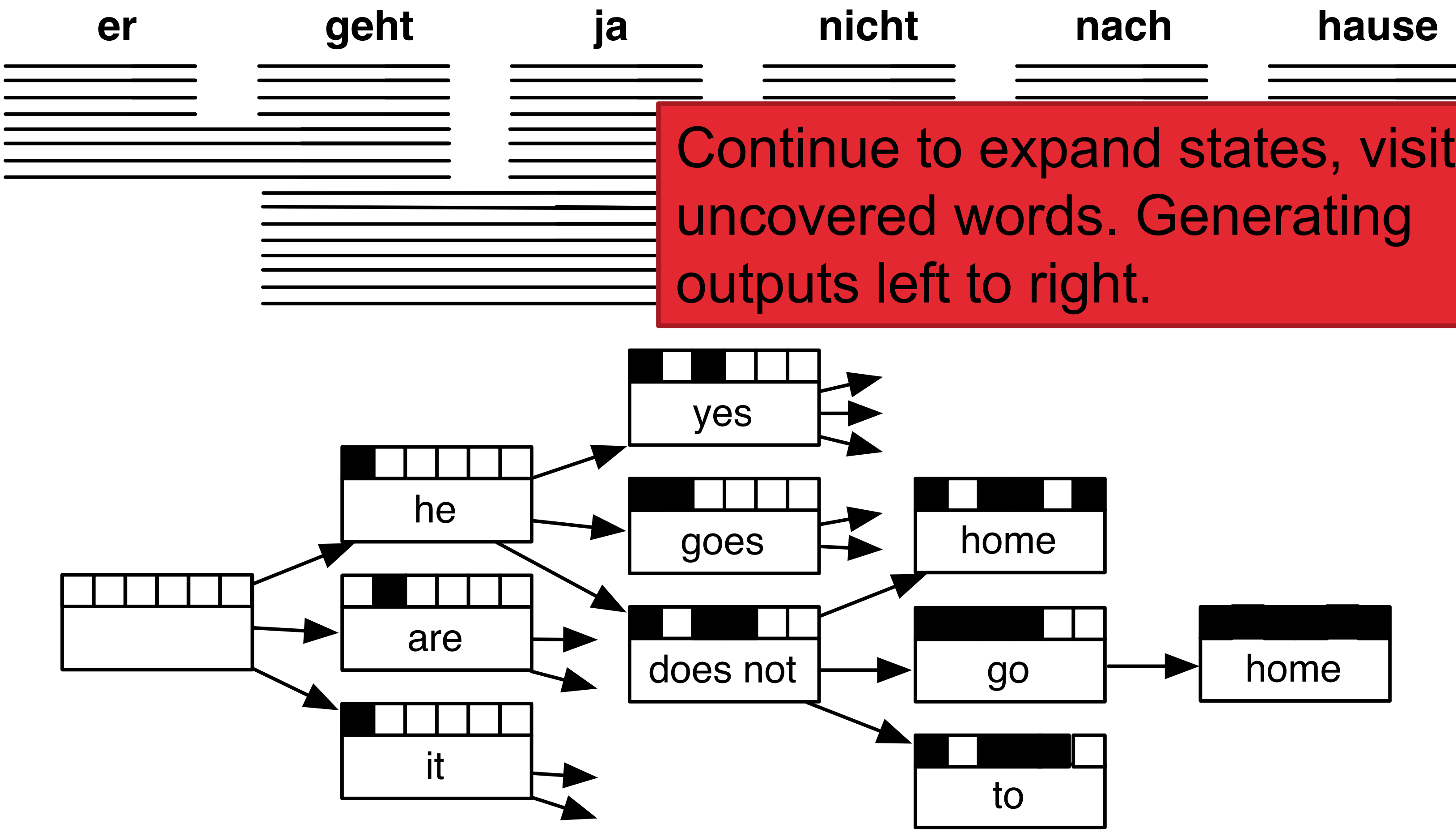




Consider all possible options to start the translation

Figure from Koehn, 2009

PHRASE-BASED DECODING



Continue to expand states, visiting uncovered words. Generating outputs left to right.

Figure from Koehn, 2009

PHRASE-BASED DECODING

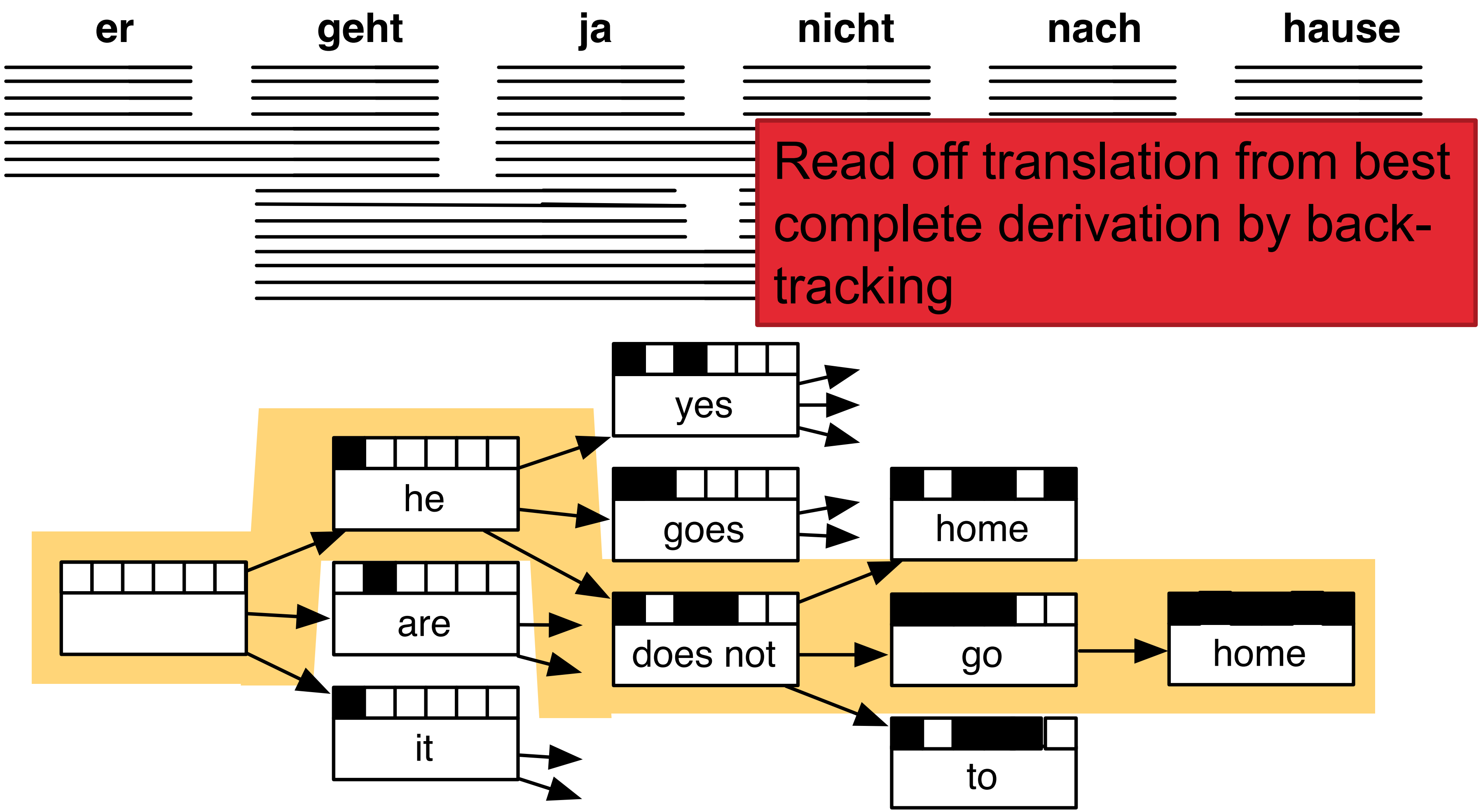


Figure from Koehn, 2009

REPRESENTING TRANSLATION STATE

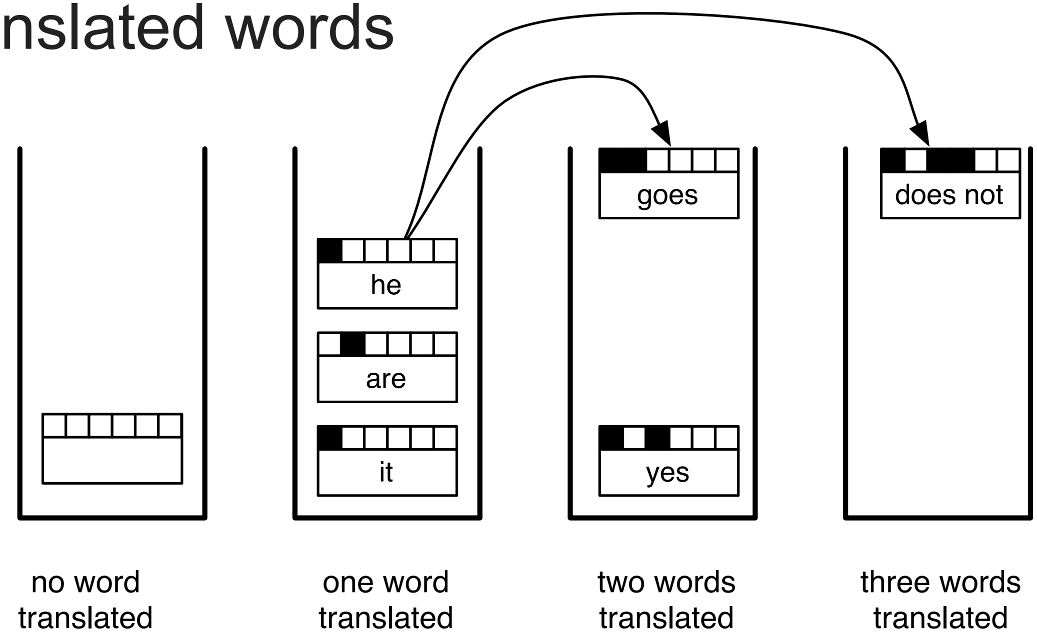
- Need to record
  - translation of phrase
  - which words are translated in bit-vector
  - last *n-1* words in E… so that *n*gram LM can compute probability of subsequent words
  - end position of the last phrase translated in the source, for scoring distortion in next step
- Together allows for the score computation to be factorised

COMPLEXITY

- Full search is intractable
  - word-based and phrase-based decoding is NP complete (Knight 99)
  - arises from arbitrary reordering
- A solution is to prune the search space
  - Use ***beam search***, a form of approximate search
  - maintaining no more than *k* options (“hypotheses")
  - pruning over translations that cover a given number of input words

LENGTH BINNING & FUTURE COST

- Each time we extend a hypothesis, store resulting translation in bin according to source coverage
  - prune each bin to no more than k entries
  - also include approximate cost of translating the untranslated words



ADVANCED EXTENSIONS

- **More Features**
  - often use many more than 3 features, although these are the central ones
  - learn to weight the effect of each feature differently (MERT)
- **Grammars and trees**
  - instead of just using phrase-pairs, can use pairs of CFG rules; parse F using one side of the translation grammar and then generate E using the other side

PHRASE-BASED MT SUMMARY

- Start with sentence-aligned parallel text
  - learn word alignments
  - extract phrase-pairs from word alignments & normalise counts
  - learn a language model
- Combine into decoding algorithm
  - … and learn feature weights
- Apply to test sentences

NEURAL MACHINE TRANSLATION

- Phrase-based approach is *rather* complicated!
- Neural approach poses question:
  - Can we throw away all this complexity, instead learn a *single model* to directly translate from source to target?
- Using deep learning of neural networks
  - learn robust representations of words and sentences
  - attempts to generate words in the target given “deep” (vector/matrix) representation of the source

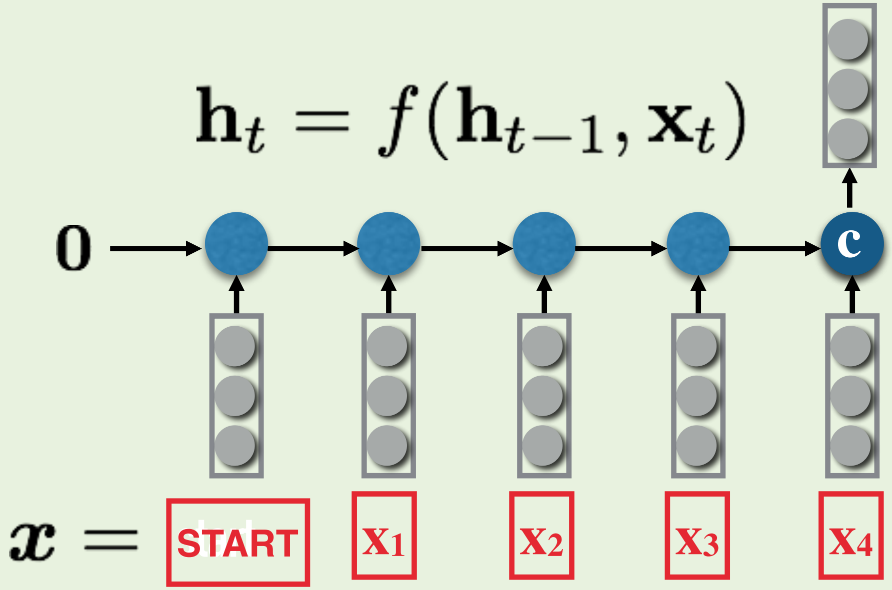
ENCODER-DECODER MODELS

- So-called "*sequence2sequence*" models combine:
  - encoder which represents the source sentence as a vector/matrix
    - akin to word2vec's method for learning word vectors
  - decoder which predicts each word in the target
    - similar to a language model, except that the decoder is conditioned on the encoder representation
- Along with lots of CPU & GPU muscle…

RECURRENT NEURAL NETWORKS (RNNS)

$$\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t)$$

$$\mathbf{c} = \mathrm{RNN}(\boldsymbol{x})$$

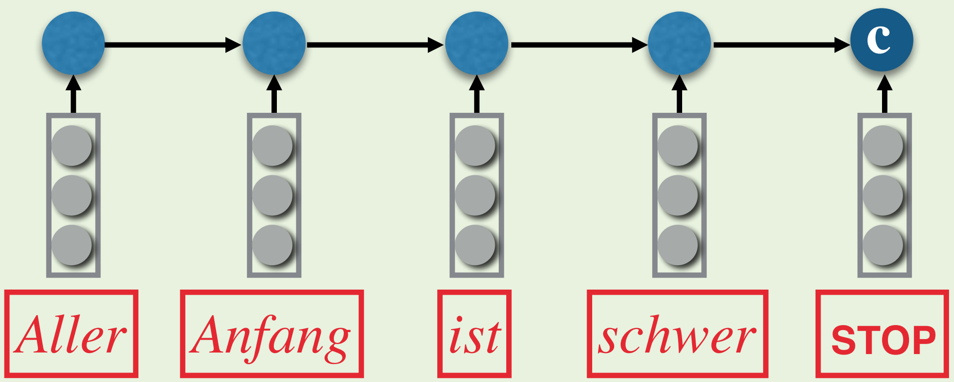


What is a vector representation of a sequence $\boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

RNN ENCODER-DECODERS

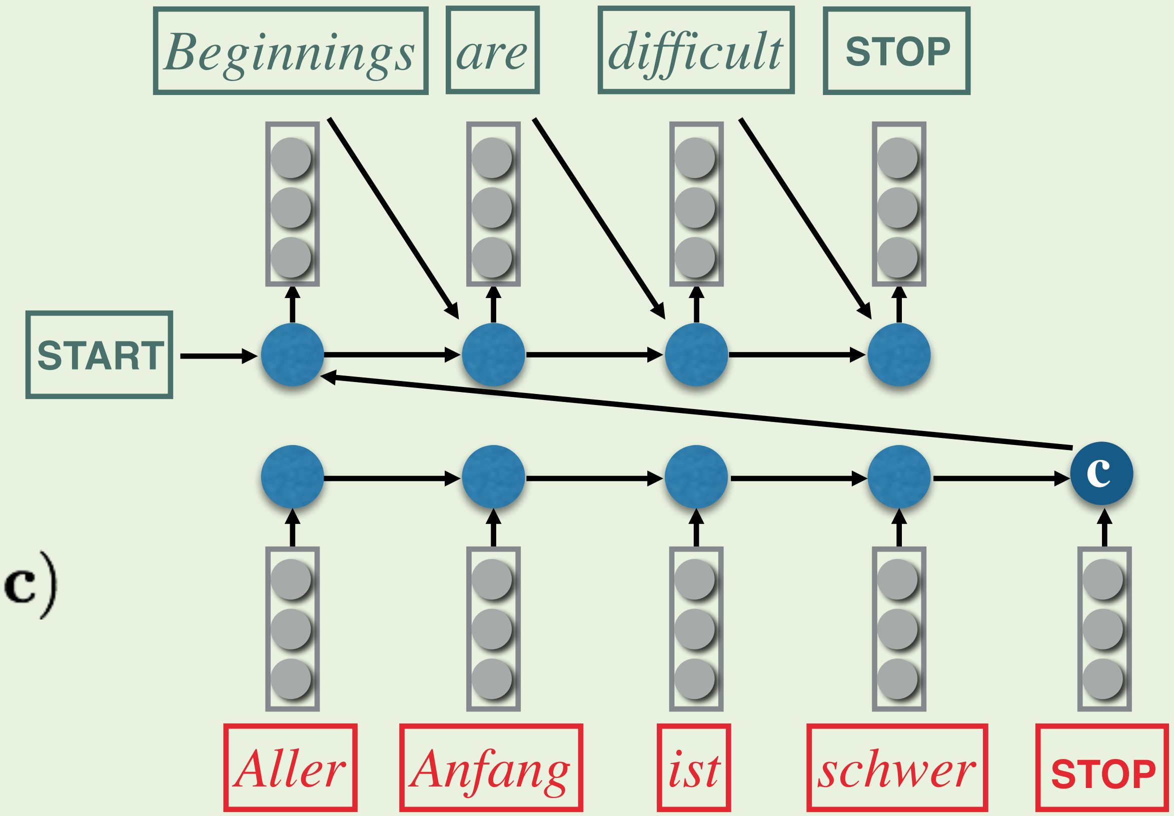
$$\mathbf{c} = \mathrm{RNN}(\boldsymbol{x})$$



What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

RNN ENCODER-DECODERS

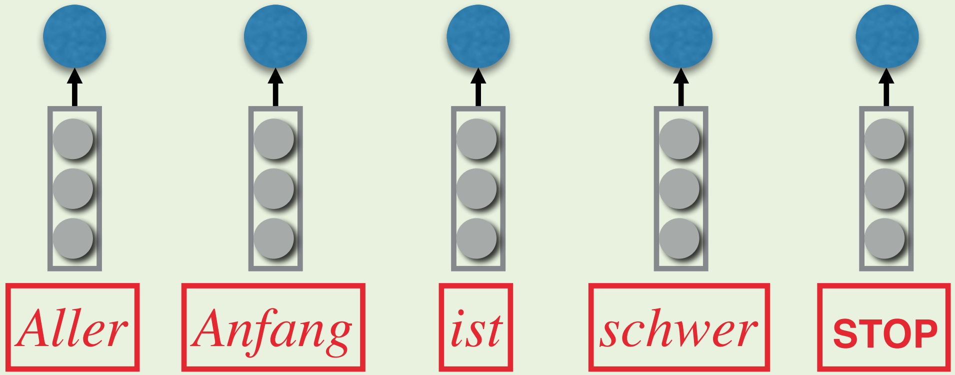


$$\mathbf{c} = \mathrm{RNN}(\boldsymbol{x})$$
$$\boldsymbol{y} \mid \mathbf{c} \sim \mathrm{RNNLM}(\mathbf{c})$$

What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

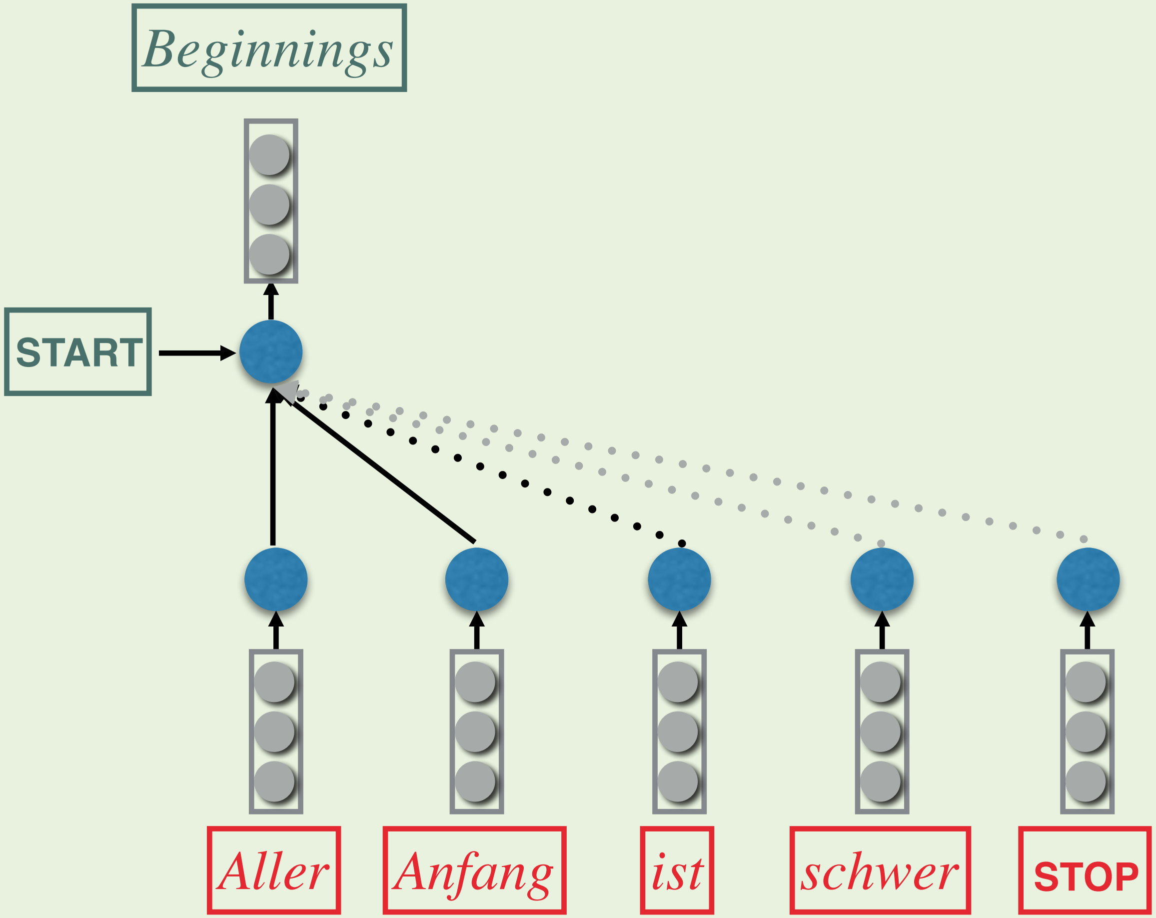
RNN ATTENTION MODEL



What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

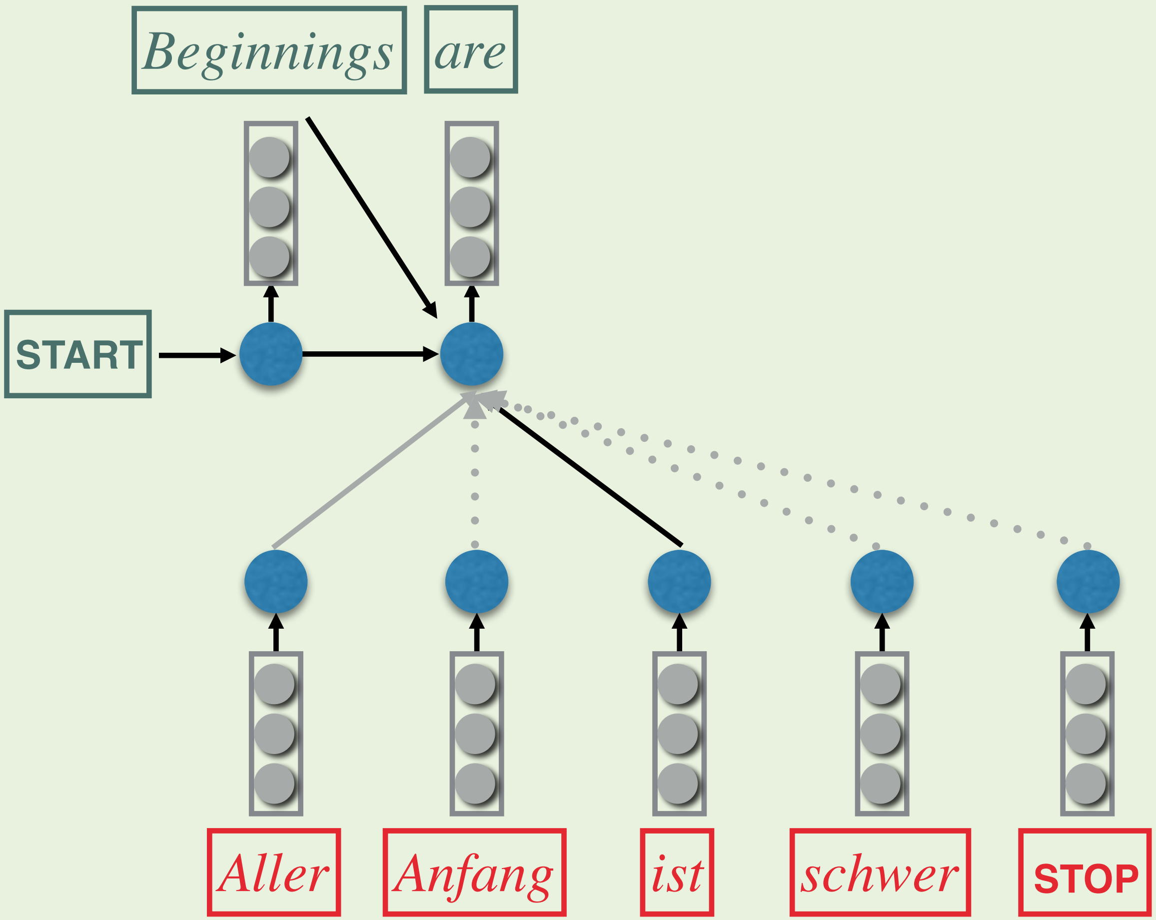
RNN ATTENTION MODEL



What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

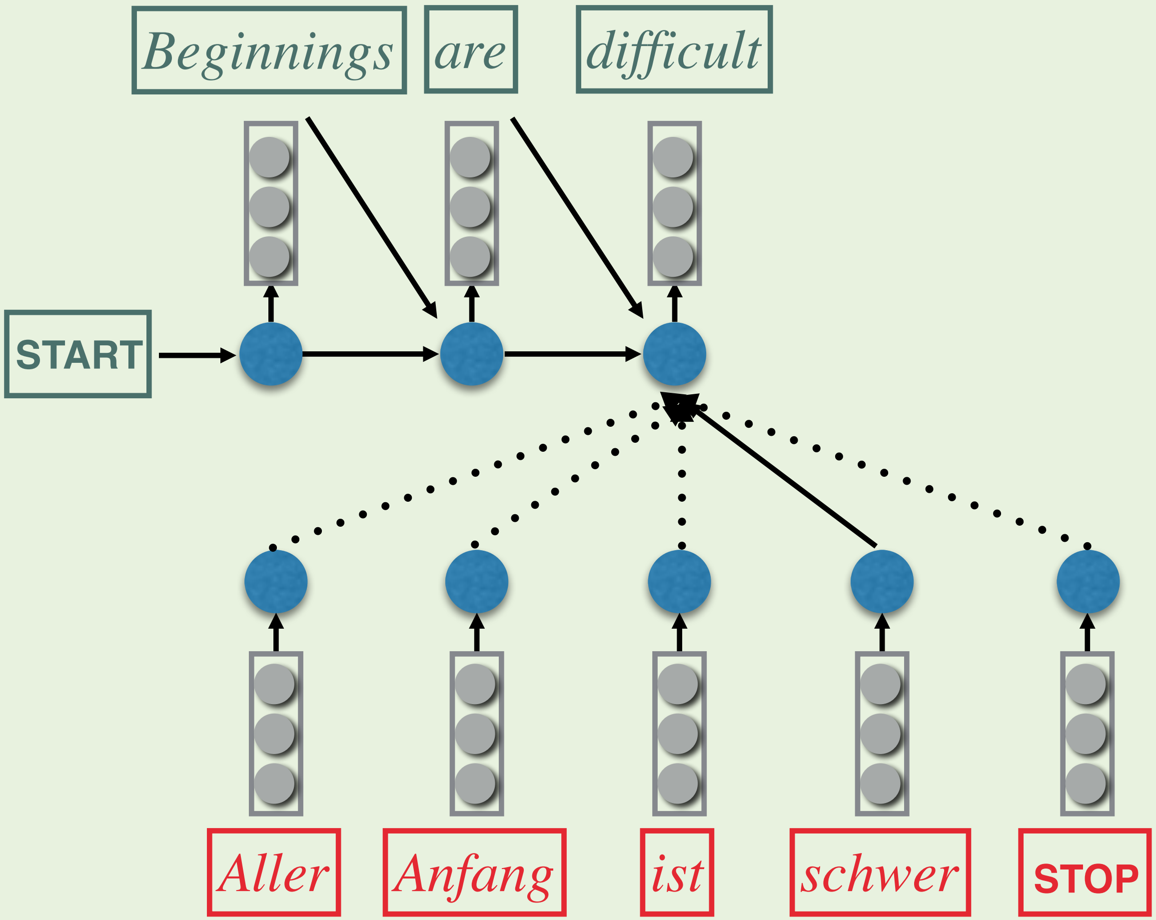
RNN ATTENTION MODEL



What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

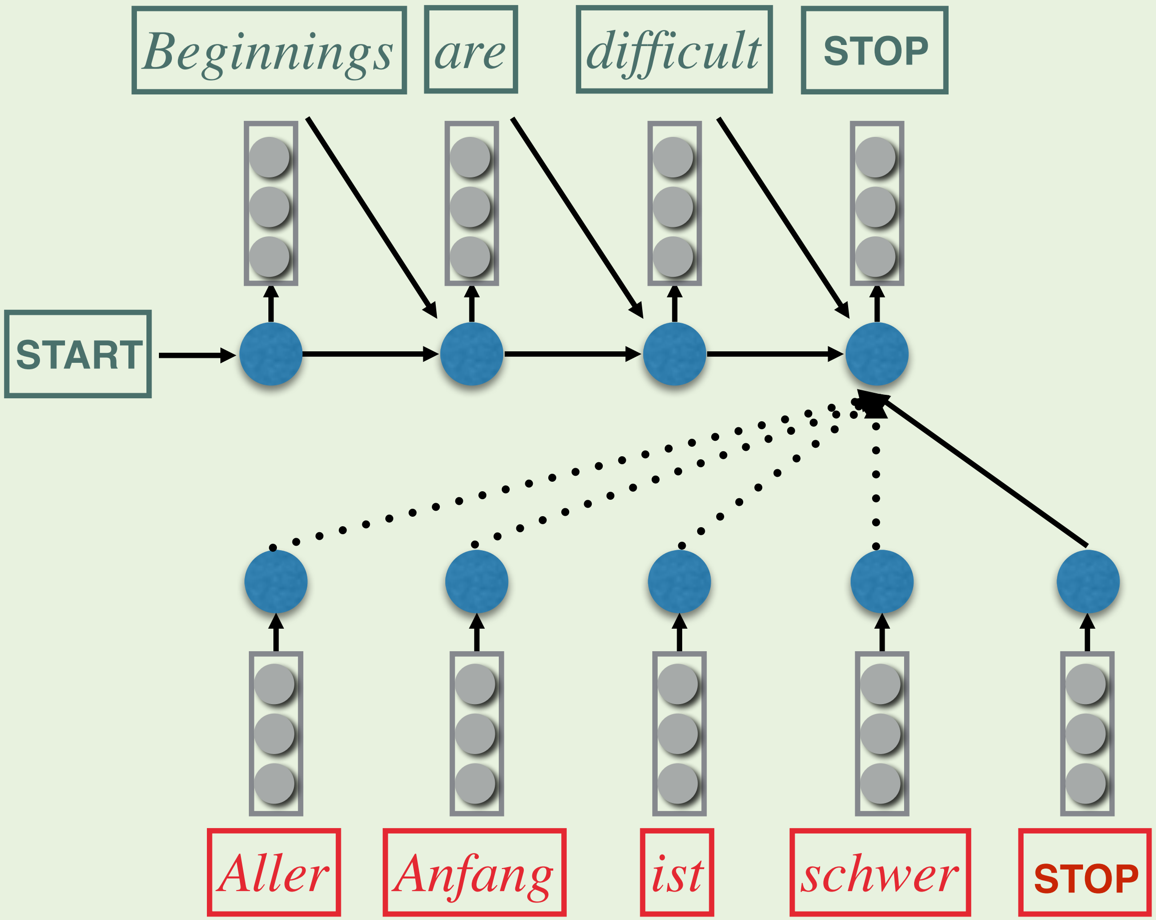
RNN ATTENTION MODEL



What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

RNN ATTENTION MODEL



What is the probability of a sequence $\boldsymbol{y} \mid \boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

APPLICATIONS OF SEQ2SEQ

- Machine translation
- Summarisation (document as input)
- Speech recognition & speech synthesis
- Image captioning & image generation
- Word morphology (over characters)
  - e.g., study → student; receive → recipient;
    play → player; pay → payer/payee
- Generating source code from text & more….

EVALUATION: DID IT WORK?

- Given input in Persian

ملبورن مهد و مرکز پیدای ش صنعت فی لمسازی و سی نما ، تلویزی ون ، رقص باله ، هنر امپرسیونی
بکهای مختلف رقص مثل نی و وگ و ملبورن شافل در استرالیا و مرکز مهم موز یک ک لاس یک و امروزی
ن ک شو

- Google translate outputs the English

ourne cradle and center of origin of the film industry and cinema, television, ballet, art,
essionism, various dance styles such as New Vogue and the Melbourne Shuffle in Australia
portant center of classical and contemporary music in this country.

- We might ask a bilingual to judge, or compare to human translation

ed to as Australia's "cultural capital" it is the birthplace of Australian impressionism, Australian rules
stralian film and television industries, and Australian contemporary dance such as the Melbourne Sh
gnised as a UNESCO City of Literature and a major centre for street art, music and theatre.

AUTOMATIC EVALUATION

- How many words are the shared between output:

lbourne cradle **and** center **of** origin **of the film** industry **and** cinema, **televis**
et**,** art**,** **impressionism**, various **dance** styles **such as** New Vogue **and the**
lbourne Shuffle in **Australia and** an important center **of** classical and
temporary **music** in this country.

- And the reference:

erred to **as Australia**'s "cultural capital" it is the birthplace of Australian
ressionism, Australian rules football**, the** Australian **film** and **television** ind
Australian contemporary **dance such as the Melbourne Shuffle**. It is recog
UNESCO City of Literature and a major **centre** for street **art, music** and the

MT EVALUATION: BLEU

- BLEU measures closeness of translation to one or more references
  - defined as:

$$\text{BLEU} = \text{bp} \times \text{prec}_{\text{1-gram}} \times \text{prec}_{\text{2-gram}} \times \text{prec}_{\text{3-gram}} \times \text{prec}_{\text{4-gram}}$$

  - weighted average of 1, 2, 3 & 4-gram precisions
    - $\text{prec}_{\text{n-gram}}$ = num *n*-grams correct / num *n*-grams predicted in output
    - numerator clipped to #occurences of *n*gram in the reference
  - and a brevity penality to hedge against short outputs
    - bp = min (1, output length / reference length)
- Shown to have fair correlation with human judgements (also many other metrics: TER, METEOR, WER, …)

SUMMARY

- Word vs phrase based MT
  - Components of phrase-base approach
  - Decoding algorithm
- Neural encoder-decoder
- Reading
  - JM2 25.7 – 25.9
  - Koehn09 5.1 – 5.2 and 6.1 – 6.2
  - JFF: **Neural Machine Translation and Sequence-to-sequence Models: A Tutorial,** Neubig 2017 https://arxiv.org/abs/1703.01619