MT: phrase based & Neural Encoder-decoder

COMP90042 Lecture 22

THE UNIVERSITY OF MELBOURNE

Overview

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

Word- and Phrase-based MT

- Seen word based models of translation
  - often 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, just adjacent words)

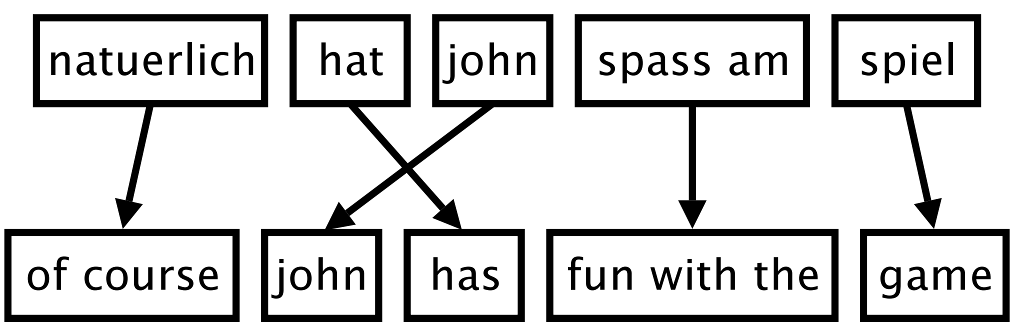
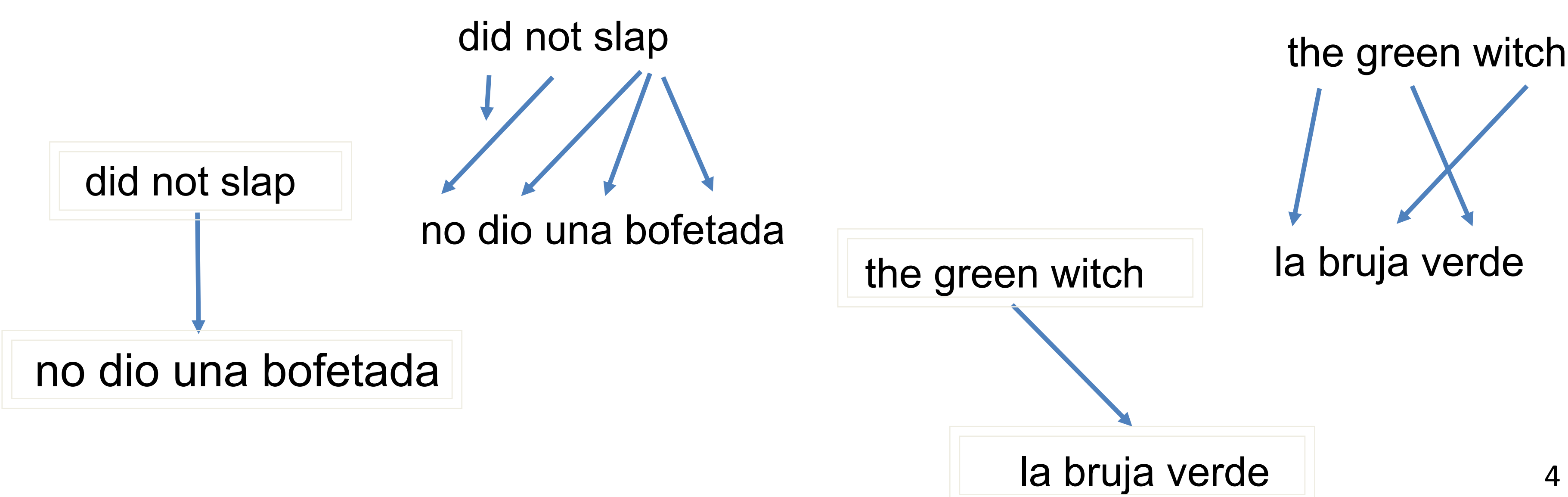


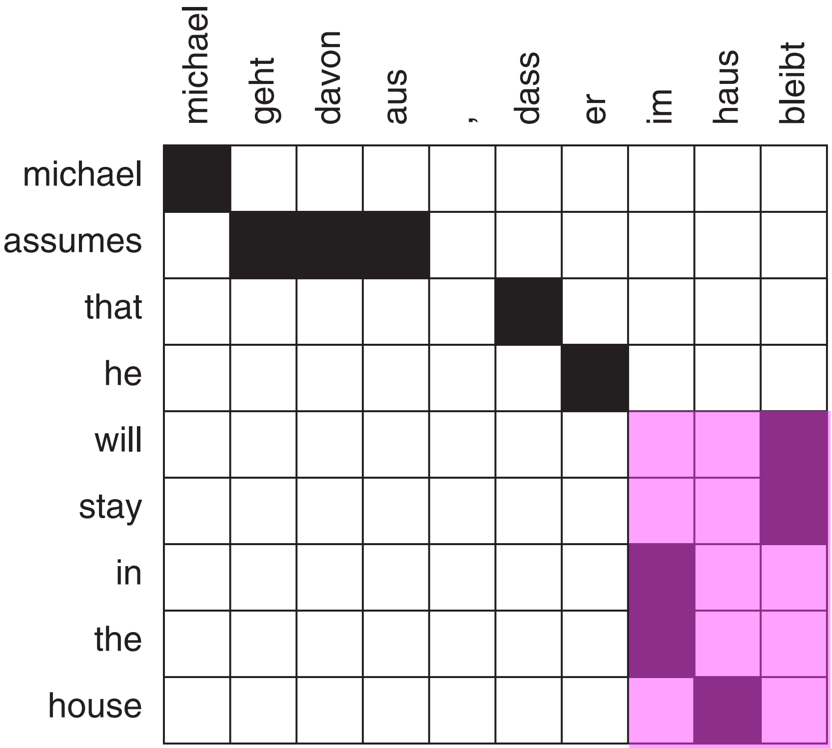
Fig from Koehn09

Why phrases not words?

- 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



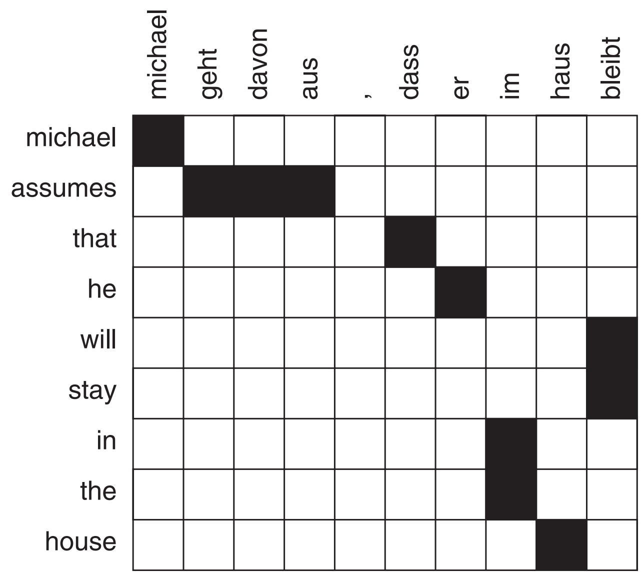
- "Extract" phrase pairs as contiguous chunks in word aligned text; then
  * compute counts over the whole corpus
  * normalise counts to produce 'probabilities'

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

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

Fig from Koehn09, Ch 5

Phrase extraction



michael – michael
michael assumes – michael geht davon aus ; michael geht davon aus ,
michael assumes that – michael geht davon aus , dass
michael assumes that he – michael geht davon aus , dass er
michael assumes that he will stay in the house
– michael geht davon aus , dass er im haus bleibt
assumes – geht davon aus ; geht davon aus ,
assumes that – geht davon aus , dass
assumes that he – geht davon aus , dass er
assumes that he will stay in the house
– geht davon aus , dass er im haus bleibt
that – dass ; , dass
that he – dass er ; , dass er
that he will stay in the house
– dass er im haus bleibt ; , dass er im haus bleibt
he – er
he will stay in the house – er im haus bleibt
will stay – bleibt
will stay in the house – im haus bleibt
in the – im
in the house – im haus
house – haus

Fig from Koehn09, Ch 5

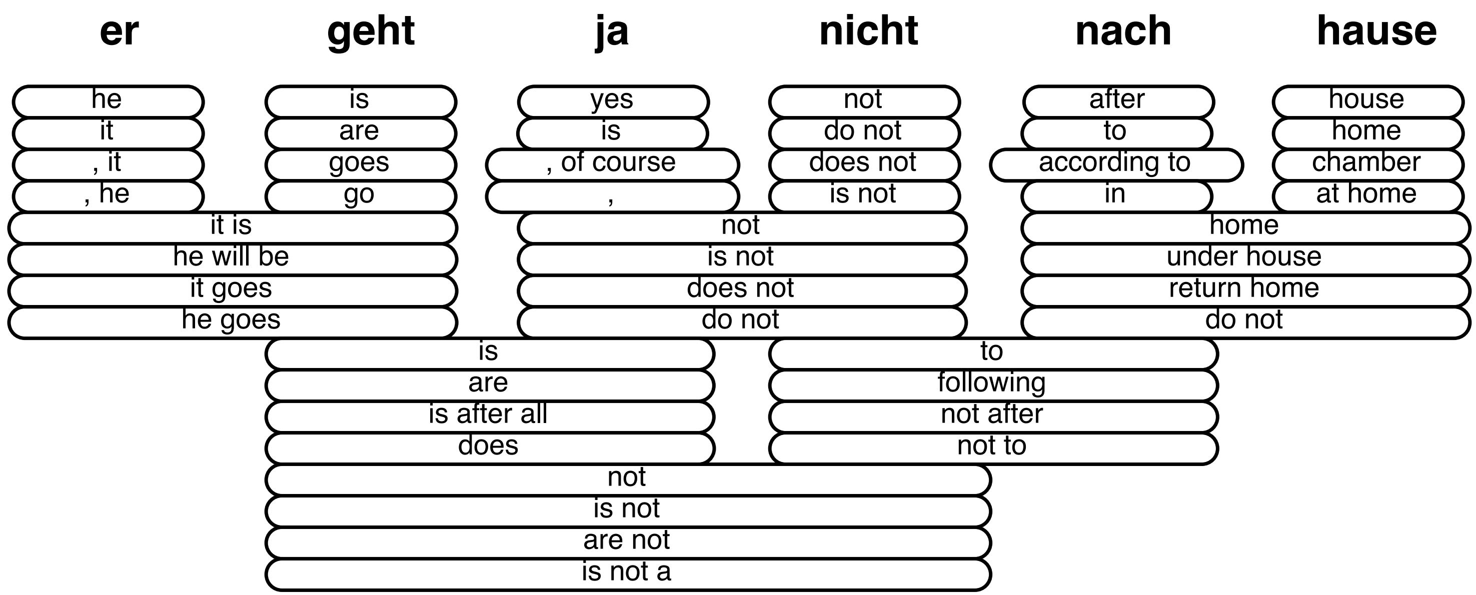
Decoding

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

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 model score, $d(start_i, end_{i-1})$, measuring amount of reordering (*minimised*) between adjacent phrase-pairs
- Search problem
  - find translation E* with the best overall score

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

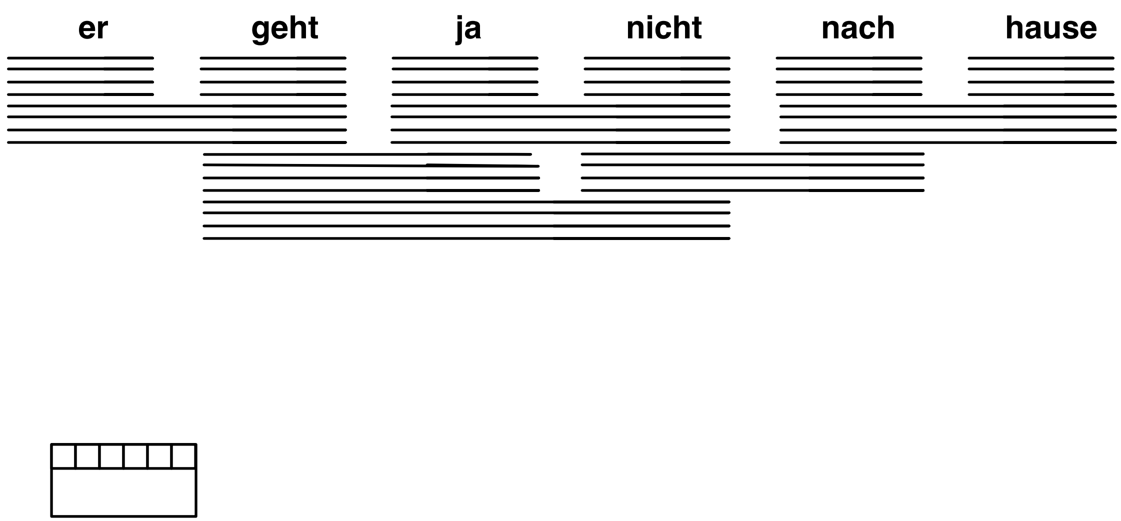
Leads to millions of possible translations…

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

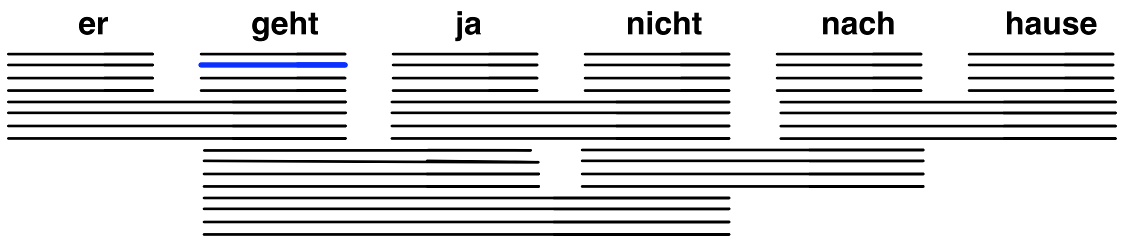
Phrase-based Decoding

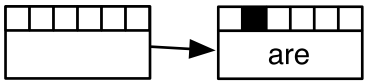


Start with empty state

Figure from Koehn, 2009

Phrase-based Decoding

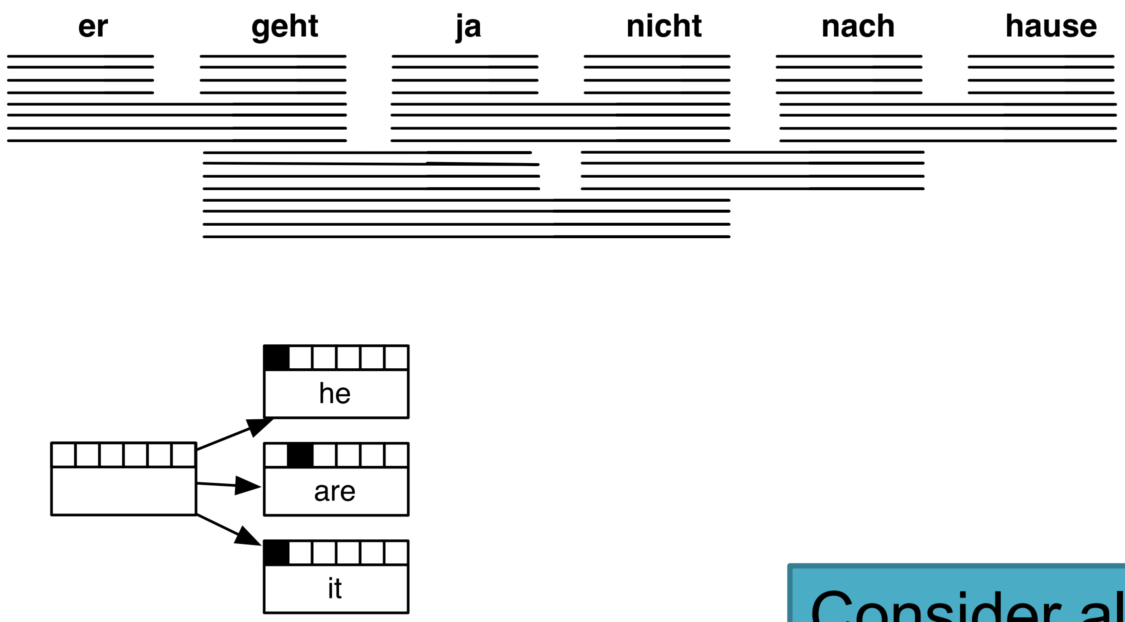




Expand by choosing input span and generating translation

Figure from Koehn, 2009

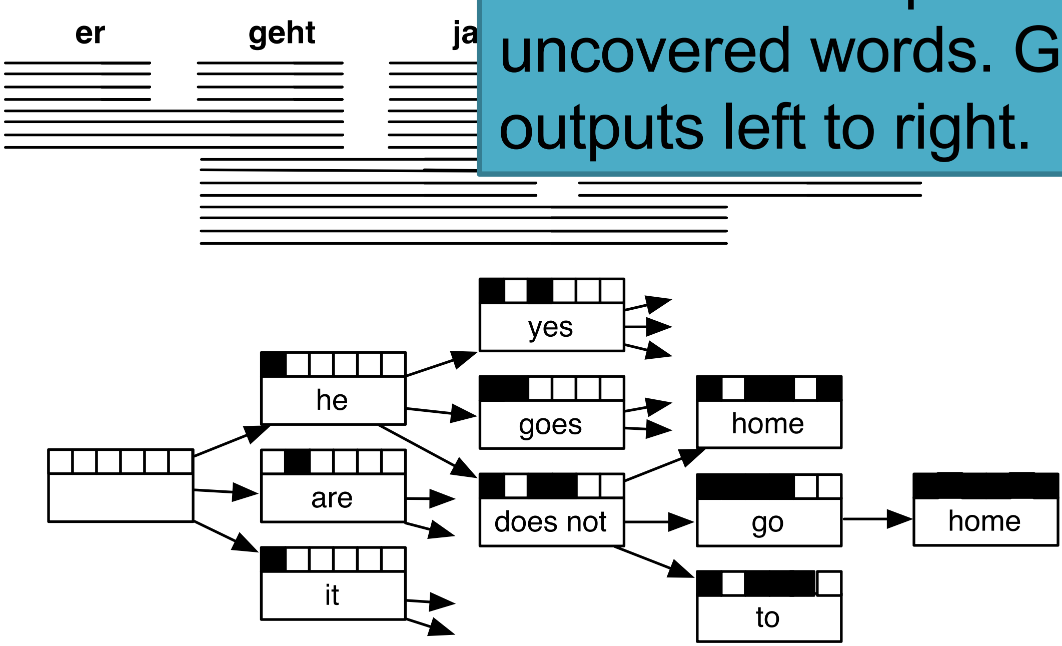
Phrase-based Decoding



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

Read off translation from best complete derivation by back-tracking

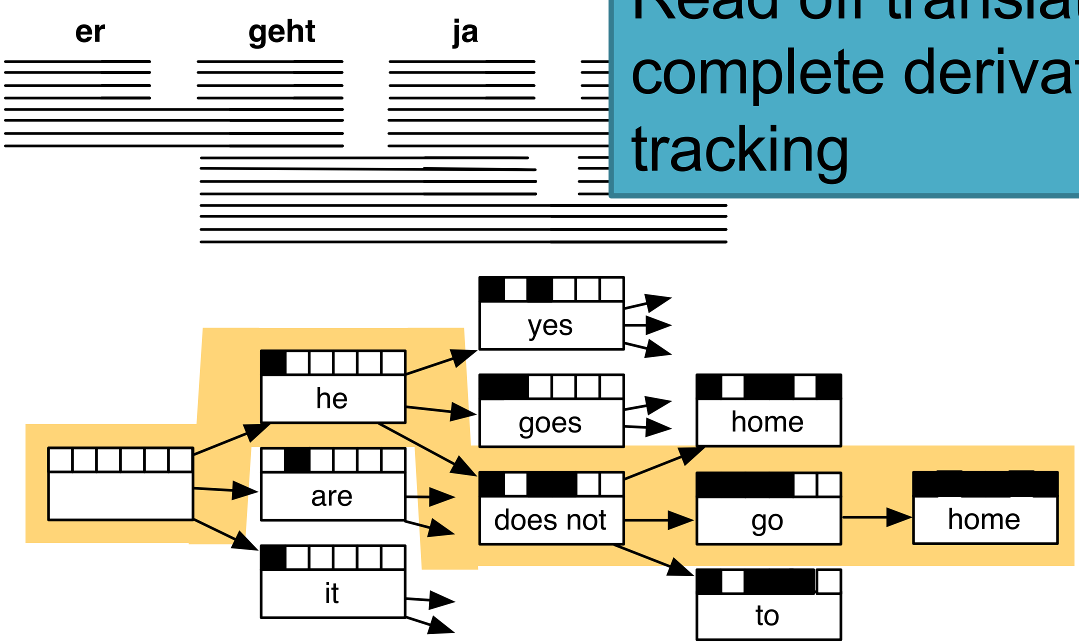


Figure from Koehn, 2009

Representing translation state

- Need to record
  - chosen translation of phrase
  - words already translated (bit-vector)
  - last *n-1* words in translation output **e**
  - end position of the last phrase translated in **f**
- Together allows for the score computation to be factorised

Complexity

- Full search is intractable
  - word-based and phrase-based decoding is NP complete — arises from allowing 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

Phrase-based MT summary

- Start with sentence-aligned parallel text
  1. learn word alignments
  2. extract phrase-pairs from word alignments & normalise counts
  3. learn a language model
- Now decode test sentences using beam-search
- State-of-the-art until 2013 *(see Moses toolkit)*

Neural Machine Translation

A type of recurrent (RNN) language model, where source sentence used as auxiliary input. Two components: *encoder* and *decoder*

Neural Machine translation

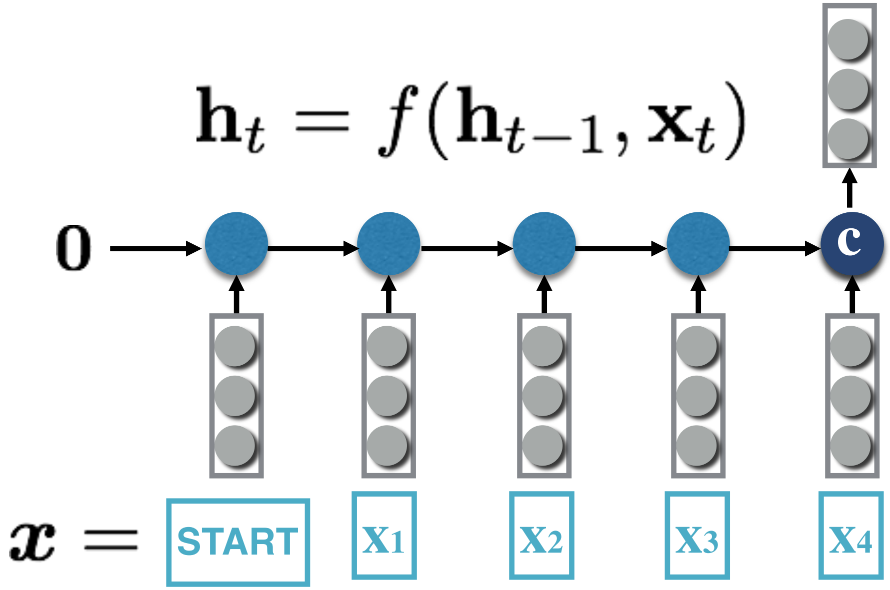
- Phrase-based approach is *rather* complicated!
- Neural approach poses question:
  - Can we instead learn a *single model* to directly translate from source to target?
- Using deep learning of neural networks
  - learn robust representations of words and sentences
  - attempts to generate words in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

Encoder-decoder models

- So-called "*sequence2sequence*" models combine:
  - encoder which represents the source sentence as a vector or matrix of real values
    - akin to word2vec's method for learning word vectors
  - decoder which predicts the word sequence in the target
    - framed as a language model

Recurrent Neural Networks (RNNs)

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

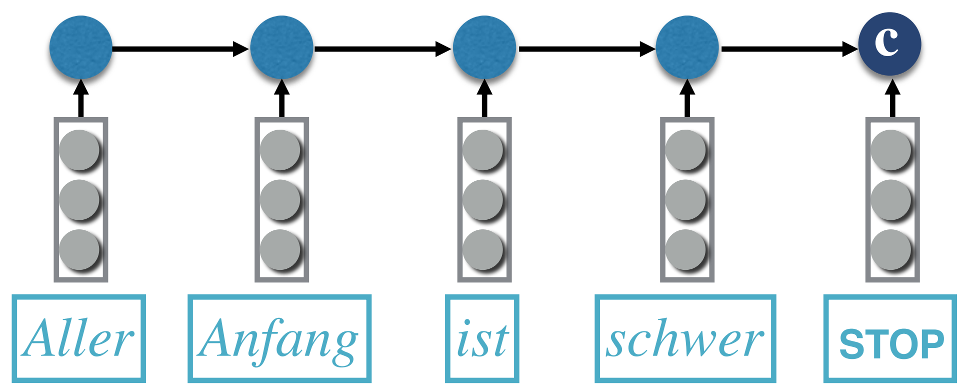


What is a vector representation of a sequence $\boldsymbol{x}$?

Slide credit: Duh, Dyer et al. 2015

RNN Encoder-Decoders

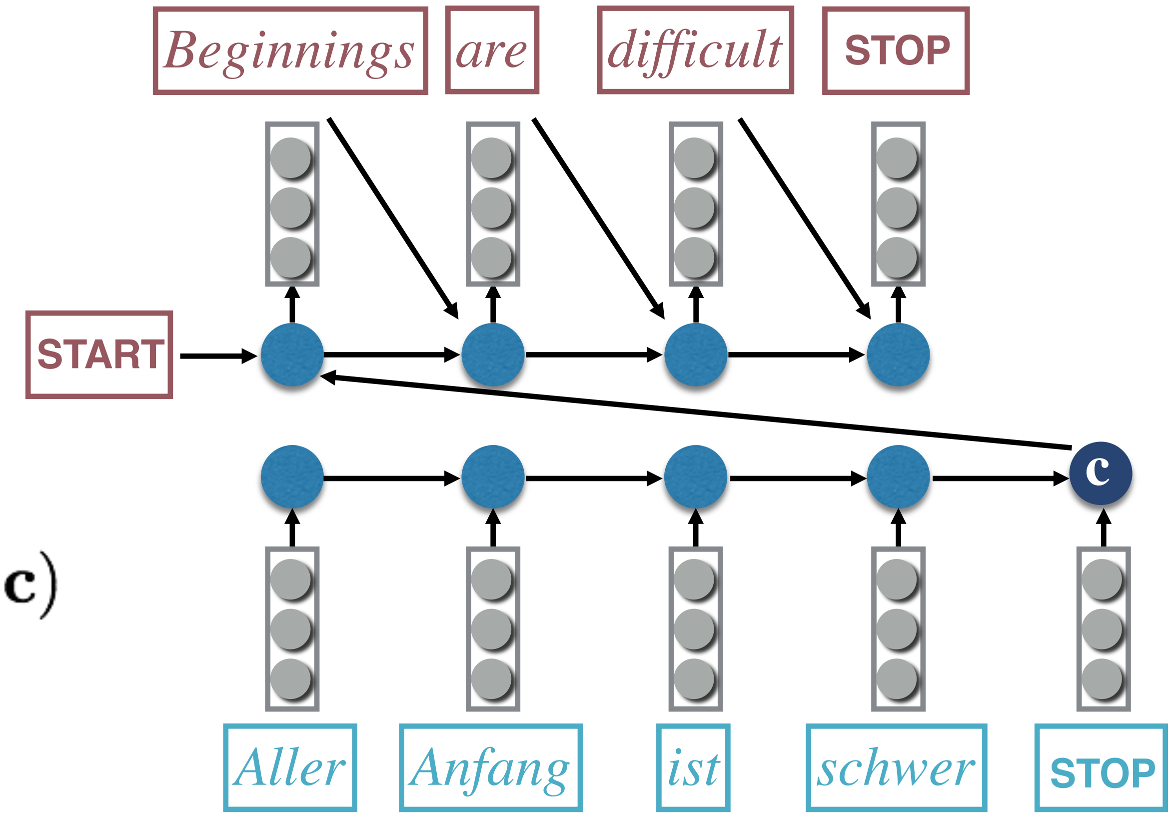
$$\mathbf{c} = \mathrm{RNN}(\boldsymbol{x})$$



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

Slide credit: Duh, Dyer et al. 2015

RNN Encoder-Decoders



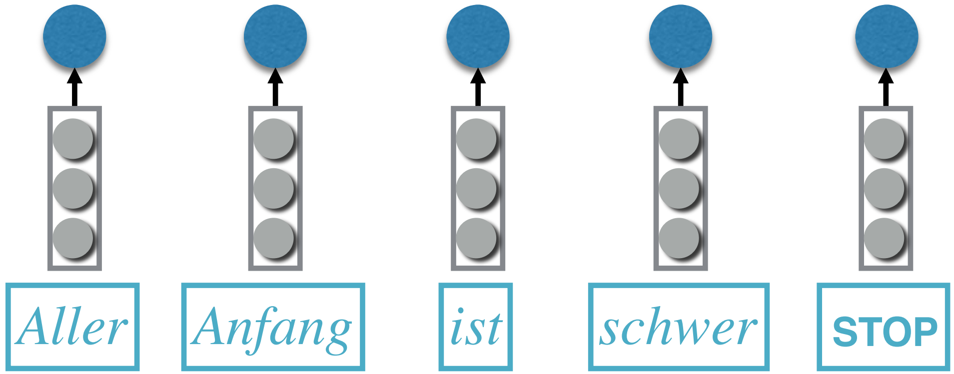
$$\mathbf{c} = \mathrm{RNN}(\boldsymbol{x})$$

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

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

Slide credit: Duh, Dyer et al. 2015

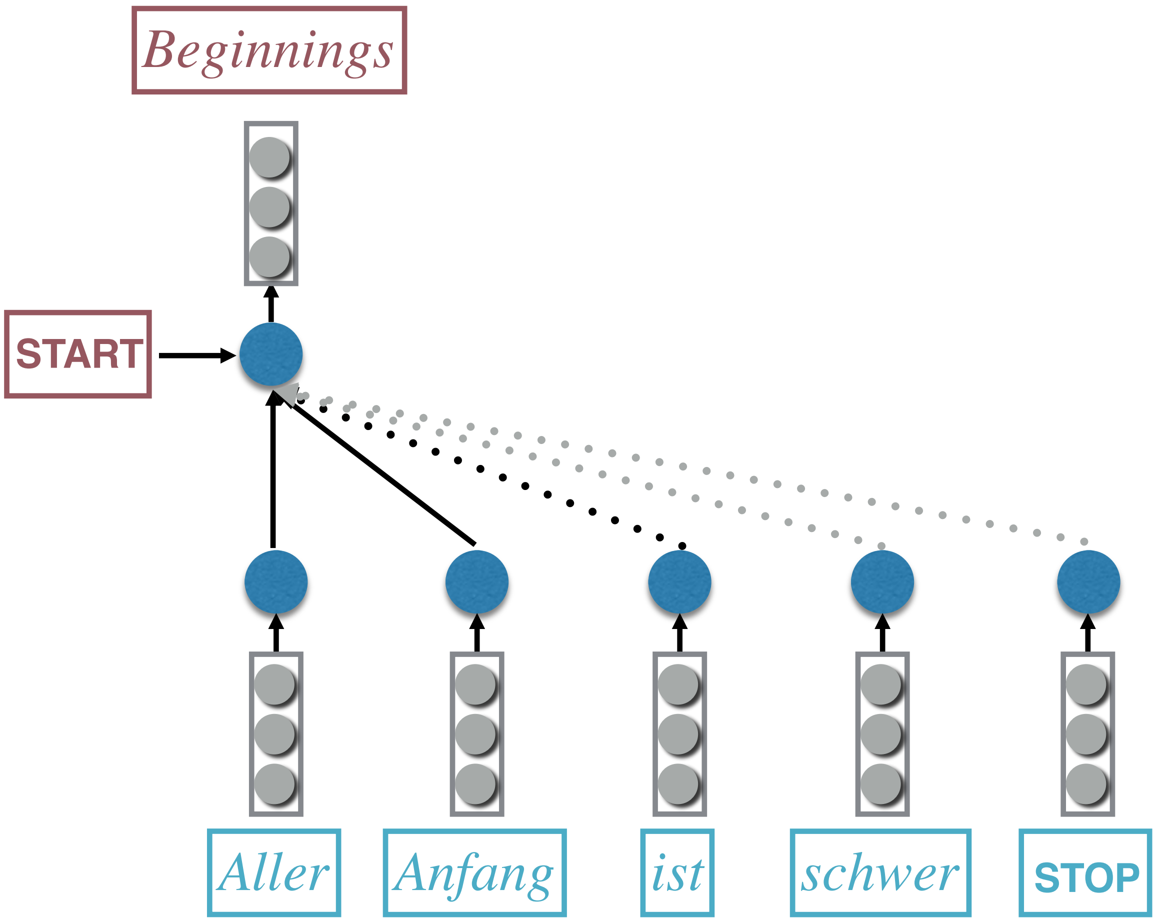
RNN Attention Model



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

Slide credit: Duh, Dyer et al. 2015

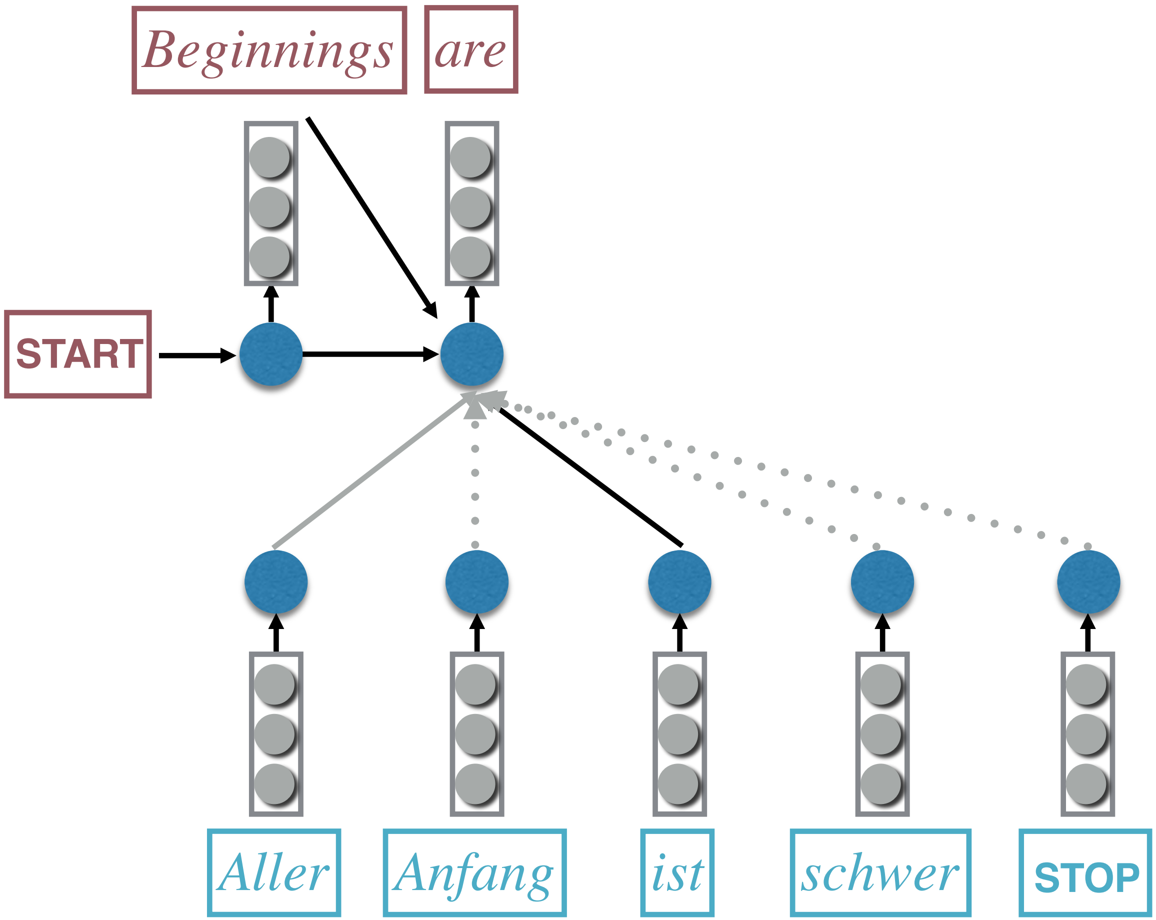
RNN Attention Model



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

Slide credit: Duh, Dyer et al. 2015

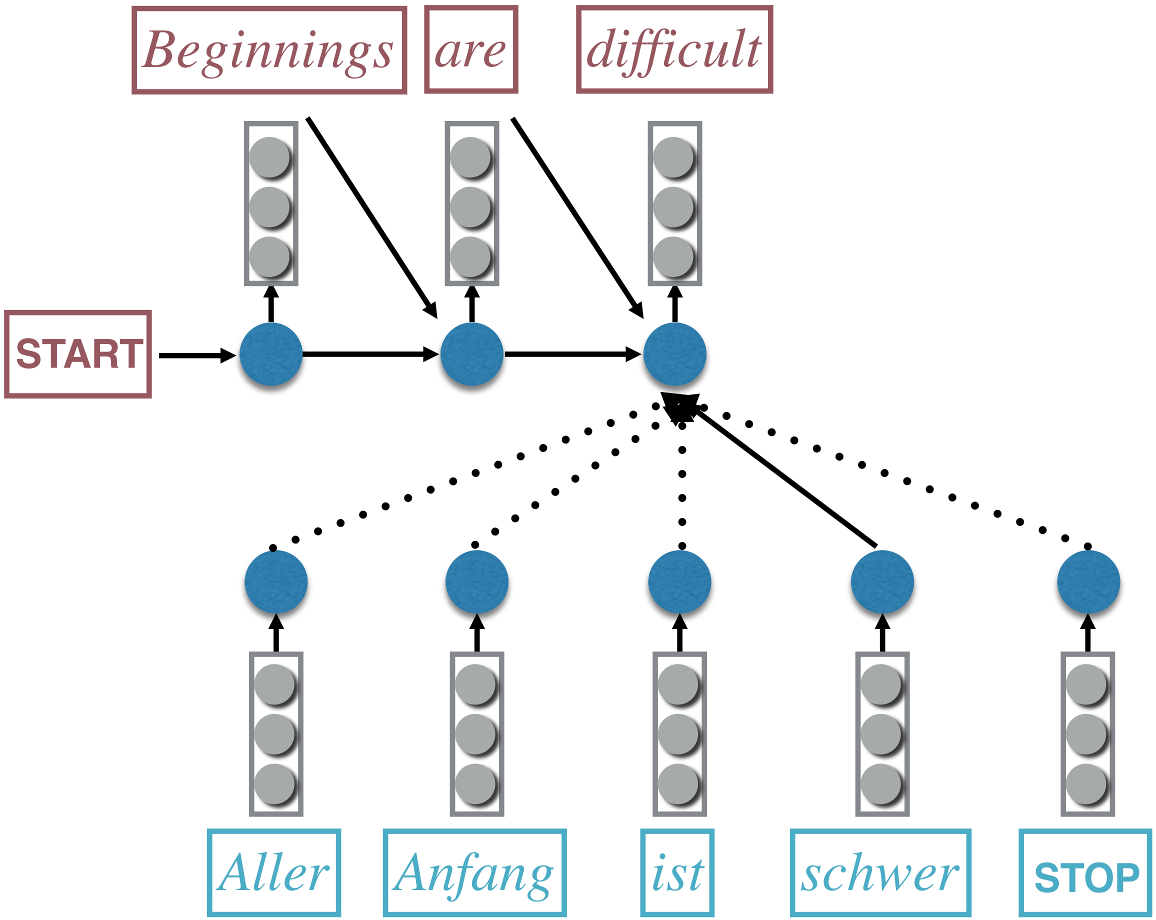
RNN Attention Model



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

Slide credit: Duh, Dyer et al. 2015

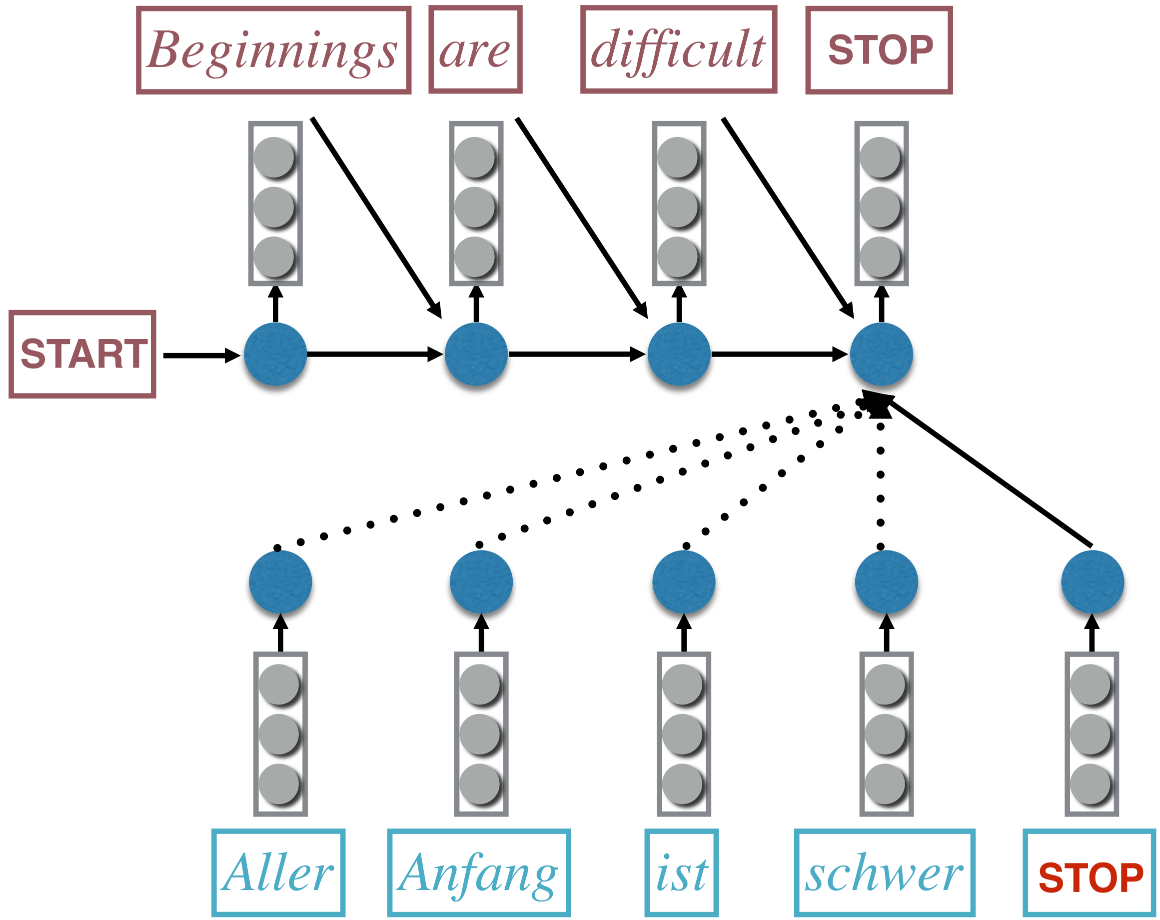
RNN Attention Model



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

Attention

- Framed as a kind of similarity function between encoder state at word *i* and decoder state at word *j*
  - learn function to take both states, and return scalar (simplest is to take the dot product)
  - normalise numbers in a softmax for a given j ("**attention**")
  - use attention to reweight encoder states to define *c*
- Context "c" now dynamic, varies during decoding

Applications of seq2seq

- Machine translation
- Summarisation (document as input)
- Speech recognition & speech synthesis
- Image captioning & image generation
- Spelling and word morphology
- Generating source code from text
- …

Evaluation

How do we know if it worked?

Evaluation: did it work?

- Given input in Persian

ملبورن مهد و مرکز پیدایش صنعت فیلمسازی و سینما ، تلویزیون ، رقص باله ، هنر امپرسیونیسم ، سبکهای مختلف رقص مثل نیو وگ و ملبورن شافل در استرالیا و مرکز مهم موزیک کلاسیک و امروزی در این کشوراست .

- Google translate outputs the English

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

- Ask bilingual to judge? Ask to rate for two components
  - **fluency:** follows grammar of English, and semantically coherent
  - **adequacy:** contains the same information as the original source document
  - or have them edit the sentence until is is adequate, and measure #changes, time spent etc (**HTER**)

Resuable evaluation

- What if we have one (or more) good translations already, e.g.

> Referred to as Australia's “cultural capital” it is the birthplace of Australian impressionism, Australian rules football, the Australian film and television industries, and Australian contemporary dance such as the Melbourne Shuffle. It is recognised as a UNESCO City of Literature and a major centre for street art, music and theatre.

- We use this text to evaluate many different MT system outputs for the same input

Automatic evaluation

- How many words are the shared between output:

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

- And the reference:

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

MT Evaluation: BLEU

- BLEU measures closeness of translation to one or more references
  - defined as: $$\text{Bleu} = \text{BP} \times \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right)$$

    a weighted average of 1 to 4-gram precisions
    - p_n = num *n*-grams correct / num *n*-grams predicted in output
    - numerator clipped to #occurrences of *n*gram in the reference
  - and a brevity penalty to hedge against short outputs
    - bp = min (1, output length / reference length)
- Correlates reasonably well with human judgements of fluency & adequacy

Summary

- Word vs phrase based MT
  - Components of phrase-base approach
  - Decoding algorithm
- Neural encoder-decoder
- Evaluation using BLEU
- Reading
  - JM2 25.7 – 25.9
  - E18 18.3 – 18.3.2 (Neural models)