AdvNLP Week 7

Machine Translation

Dr Julie Weeds, Spring 2024

Previously

- Distributional semantics
- Language models
- Neural language models
- Sequence labelling
- Sequence classification

Overview

- What makes machine translation (MT) hard?
- Evaluation of MT
- Classical MT (Pre 1990s)
- Statistical MT (1990-2015)
  - Word-based models
  - Phrase-based models
- Neural MT (2015 -)
  - Encoder-decoder models

… was hard!

French	English (Google 2014)
Pour une journée au ski réussie à Samoëns, n'hésitez pas, consultez les prévisions météo sur la station, les conditions d'enneigement et l'ouverture des pistes. Profitez d'une bonne journée au ski à Samoens, si vous venez en week end ou pour une semaine, bénéficiez de prevision météo à 2 jours sur notre site et à plus long terme sur le site de Météo Chamonix !	For a successful day at Ski Samoens , please, check the forecast weather station on the snow conditions and opening tracks. Enjoy a good day skiing Samoens , if you come on a weekend weekend or for a week, get forecasting weather two days on our website and in the longer term on the site Weather Chamonix !

… was hard!

French	English (website/human)
Pour une journée au ski réussie à Samoëns, n'hésitez pas, consultez les prévisions météo sur la station, les conditions d'enneigement et l'ouverture des pistes. Profitez d'une bonne journée au ski à Samoens, si vous venez en week end ou pour une semaine, bénéficiez de prevision météo à 2 jours sur notre site et à plus long terme sur le site de Météo Chamonix !	To ensure a great ski day in Samoëns, don't hesitate to check the weather report of this ski station, the snow report and the slopes opening. Enjoy a great ski day in Samoëns, if you come for the week-end or for a week, check the Samoëns weather report for the next 2 days on this website and for a longer period of time on the Météo Chamonix website.

… is solved?

French	English (Google 2019)
Pour une journée au ski réussie à Samoëns, n'hésitez pas, consultez les prévisions météo sur la station, les conditions d'enneigement et l'ouverture des pistes. Profitez d'une bonne journée au ski à Samoens, si vous venez en week end ou pour une semaine, bénéficiez de prevision météo à 2 jours sur notre site et à plus long terme sur le site de Météo Chamonix !	For a successful ski day in Samoens, do not hesitate, check out the weather forecast for the resort, the snow conditions and the opening of the slopes. Enjoy a good day skiing in Samoens, if you come for a weekend or for a week, enjoy 2-day weather forecast on our site and longer term on the site of Weather Chamonix!

Why is/was MT hard?

- Lexical differences
- Structural differences (morphological differences and syntactic differences)
- Study of systematic cross-linguistic similarities and differences is called **linguistic typology**
  - See World Atlas of Language Structures (Dryer and Haspelmath, 2013)

Lexical Divergences

- Homonymy and polysemy (in both languages)
- e.g., "I *know* that machine translation is hard."→ French *savoir*
- whereas "I *know* David Weir." → French *connaître*
- different distinctions in different languages e.g., Chinese distinguishes *older brother* and *younger brother*

I met my brother → 我遇到了我的哥哥
Wǒ yù dàole wǒ dí gēgē

Lexical Divergences

- can be grammatical e.g., the English verb *to like* corresponds to the German adverb *gern*
- gender marking
  - *they* -> *elles* or *ils* in French?
  - In Spanish, different forms for male and female animals)
- lexical gaps
  - la baguette
  - le weekend

Morphological differences

unhappiness

=un + happy + ness

- *agglutinative* vs *fusion*

  agglutinative => different morphemes within a word are clearly differentiable, e.g., Finnish

  fusion => morphemes not clearly distinguishable, may contain multiple bits of information e.g., Indo-European languages

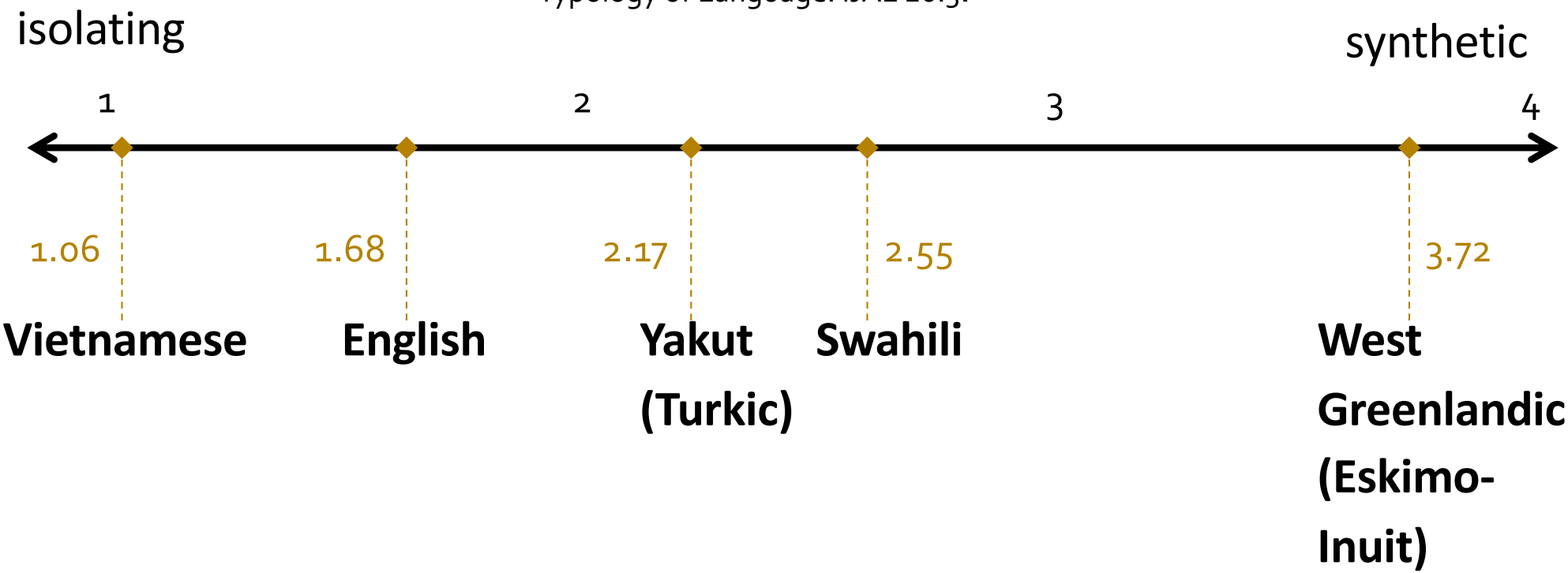
- isolating vs polysynthetic

  isolating => little or no morphological change e.g., Chinese

  polysynthetic => high morpheme to word ratio, the ability to form words which are equivalent to sentences in other languages, e.g., many Amerindian languages

Morphemes per Word

Joseph Greenberg. 1954. A Quantitative Approach to the Morphological Typology of Language. IJAL 26:3.



Many Morphemes per word: Turkish

uygarlaştıramadıklarımızdanmışsınızcasına

uygar+laş+tır+ama+dık+lar+ımız+dan+mış+sınız+casına

Behaving as if you are among those whom we could not cause to become civilized

Syntactic Differences

- See Jurafsky and Martin Chapter 13
  - SVO vs SOV vs VSO
  - adjective noun order
  - prepositions vs postpositions
  - head marking vs dependent marking
  - verb-framed vs satellite framed
  - pro-drop languages, referential density, hot and cold languages
  - idiosyncratic constructions e.g., “There burst into the room three men with guns.”

Evaluation

- Human Raters
  - Fluency :

    ratings on scale of 1 -5

    Cloze task

    delete every nth word – can human readers guess the missing words?
  - Fidelity

    adequacy/informativeness

    bilingual vs monolingual raters
  - post-edit cost
- Automatic Evaluation e.g., BLEU, chrF

BLEU

Machine: check the forecast weather station

Human 1: consult the resort weather forecast

Human 2:check the weather forecast at the resort

- Average Unigram Precision= ½ (3/5 + 4/5)
- What are the flaws?
- Why not recall?

Modified precision

Machine: the the the the the

Human 1: consult the resort weather forecast

Human 2:check the weather forecast at the resort

- average unigram precision = ½ (5/5 + 5/5)=1
- is this a good translation?

Modified precision

- For each word in the machine translation, take the maximum number of times it occurs in any human reference
- For example, m_{max}(the) = 2
- Restrict the number of times a word can appear in machine translation to its m_{max}
- In above example, modified unigram precision = 2/5

BLEU (continued)

- computes modified precision for unigrams, bigrams, trigrams and often quadrigrams
- combines using geometric mean
- incorporates a penalty for translations which are too short
- good for evaluation of incremental changes to same general architecture
- see Papineni 2002

chrF

- Character-based F score(Popovic 2015)
- Overcomes problems with different tokenization standards
- In practice, if there are different possible human translations, machine translation cannot capture all of the variation, so recall may be lower than precision
  - But this is the same for all of the hypotheses (different machine translations)
  - So, recall still provides a robust way of balancing precision

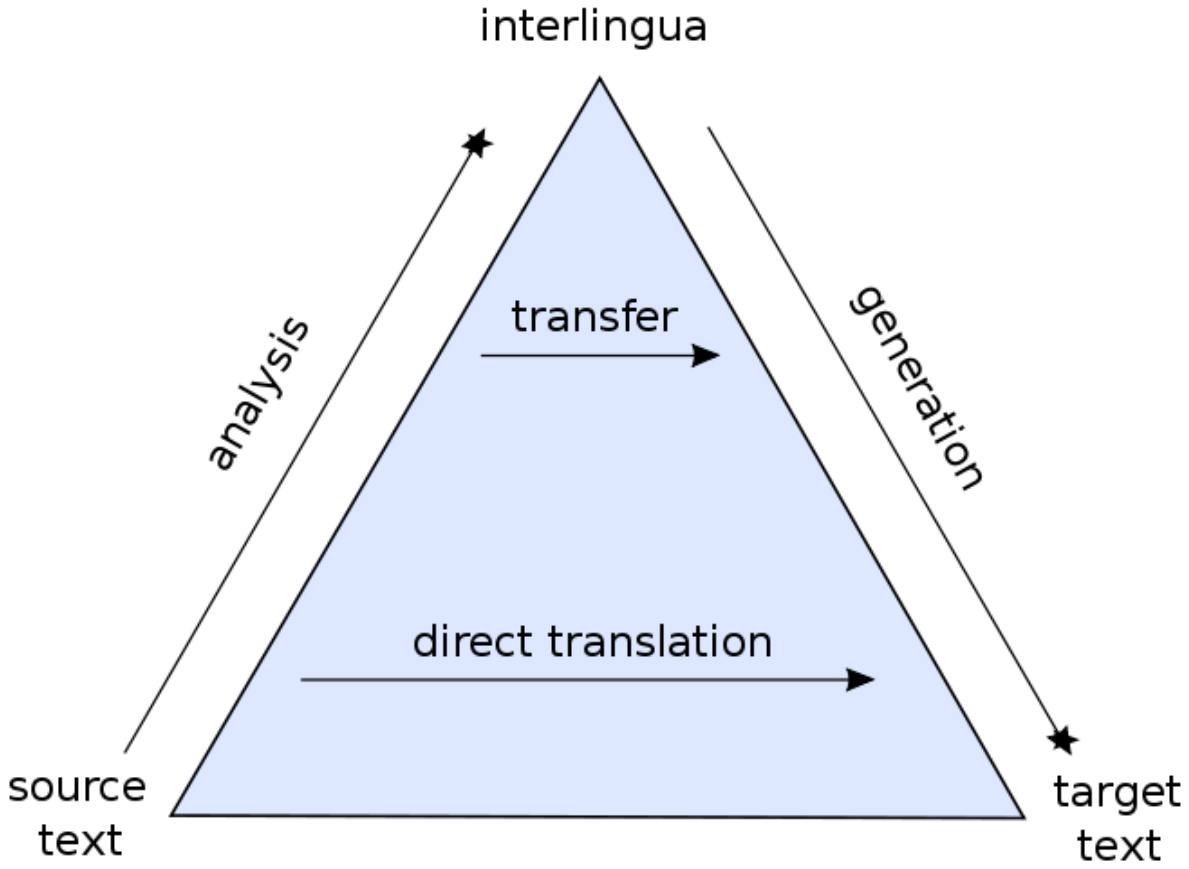
chrF (cont)

- chrP = percentage of character 1-grams, 2-grams, … , k-grams in the hypothesis that occur in the reference, averaged
- chrR = percentage of character 1-grams, 2-grams, …, k-grams in the reference that occur in the hypothesis, averaged

$$chrF\beta = (1+\beta)^2 \frac{chrP \,.\, chrR}{\beta^2 . chrP + chrR}$$

β=2 gives twice as much weight to chrR as to chrP

Classical MT: Vauquois Triangle



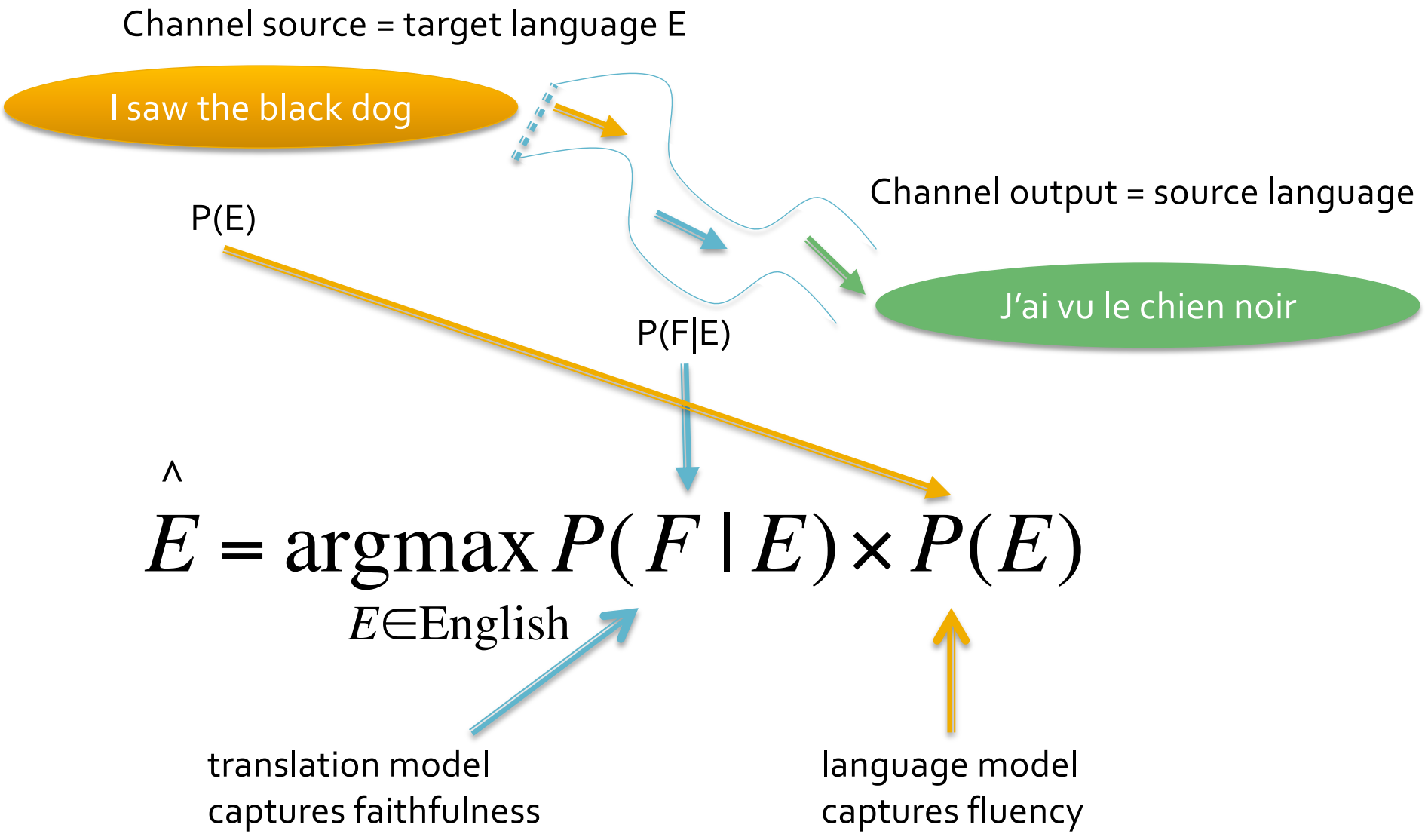
Classical MT systems developed before 1990s were generally **rule-based** systems used 3 basic approaches:

- Direct translation
- Transfer-based
- Interlingua based

Statistical MT

- focus on the result NOT the process
- What does it mean for a sentence to be a translation of some other sentence?
- faithfulness and fluency
- based on probabilities derived from *parallel corpora*
  - *Sentences in a source language matched with sentences in the target language*

Bayesian Noisy Channel Model



P(E): fluency

- The language model captures the fact that

  I the dog black saw.

  is less fluent / less probable in English than

  I saw the black dog.

P(E) can be obtained from any monolingual corpus

- Most systems used an n-gram language model e.g., a bigram model

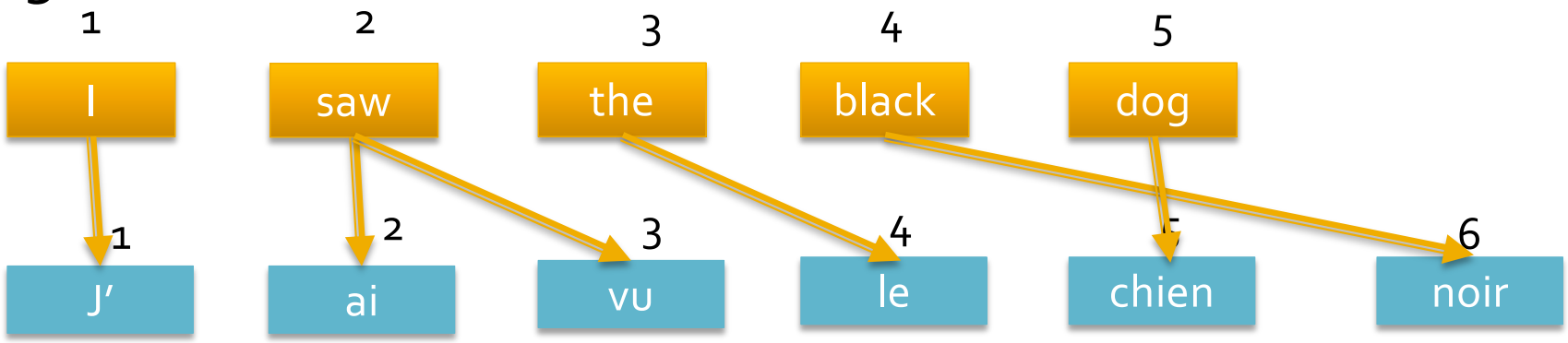
$P(E_1) = P(I \mid START) \times P(the|I) \times P(dog|the) \times P(black|dog) \times P(saw|black) \times P(END|saw)$

$P(E_2) = P(I \mid START) \times P(saw|I) \times P(the|saw) \times P(black|the) \times P(dog|black) \times P(END|dog)$

- More sophisticated models (e.g., based on grammar) could be used

P(F|E): faithfulness

- Simplest translation models are based on *word alignment*
- A word alignment is a mapping between words in F and words in E
- Common simplifying assumption is that it is a one-to-many alignment – each French word comes from exactly one English word

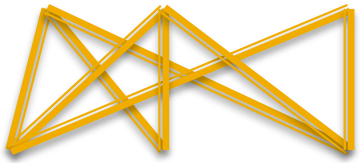


Alignment can be represented by a vector of indices.
Here, the vector would be [1, 2, 2, 3, 5, 4]

Estimating Translation Probabilities

- Parallel corpora → sentence-aligned data

…le chien noir …



…the black dog ..

- English and French words but no connections between them
- How can we use this to estimate the probability of a French word being generated by an English word?

Example

The three goats all loved to eat grass.
Les trois chèvres ont tous adoré manger de l'herbe.
They ate grass all day long on the hill.
Ils mangeaient de l'herbe toute la journée sur la colline.
But they never crossed the bridge to eat the grass on the other side.
Mais ils n'ont jamais traversé le pont de manger l'herbe de l'autre côté.
They never crossed the bridge because the Troll lived under the bridge.
Ils n'ont jamais traversé le pont parce que le Troll vivait sous le pont.
The Troll was very bad.
Le Troll était très mauvaise.
He ate anyone who dared to cross his bridge.
Il a mangé quiconque osait traverser son pont.

Expectation Maximization (EM)

- initialise model parameters (all connections equally likely)
- assign probabilities (E-step)
- estimate parameters (M-step)
- iterate

From word-based models to phrase-based models

- Word-based models assume one-to-many alignment of words
- Word-based models cannot (easily) handle non-compositional phrases
- Phrase-based models treat phrases as atomic units
- Many-to-many translation can handle non-compositional phrases
- Was (until 2016) the state-of-the-art used by Google Translate.

Phrase-alignment

- Generated by running word alignment algorithms in both directions to give
  - a one-to-many alignment
  - a many-to-one alignment
- Classifiers developed to decide how to symmetrize the alignments somewhere between intersection and union

Phrase Translation Table

- Given a phrase alignment, we can store each pair of aligned phrases in a phrase translation table together with its MLE translation probability:

$$\phi(f,e) = \frac{\text{count}(f,e)}{\sum_{f} \text{count}(f,e)}$$

- This gives us the ``translation options'' for each phrase at decode time.

Standard Model for PBMT

- For translating French (source language) to English (target language), use a log-linear model:

$$\hat{E} = \underset{E \in English}{\operatorname{argmax}} \; P(E|F) = exp\left(\sum_i \lambda_i h_i(e, f)\right)$$

- The feature functions h_i are typically
  - a language model;
  - a reordering model;
  - a word penalty; and
  - various translation models (phrase translation and word translation)
- Discriminative or generative? Why?

Decoding for Phrase-Based Statistical MT

- Finding the sentence which maximises translation probability is a search problem
- Exhaustive search impossible!
- Decoders in MT tend to use best-first search
- Maintain a priority queue (or stack) with all partial translation hypotheses and their scores
- Use the phrase-translation table to limit the search space to target language sentences which are possible translations of the source language sentence
- Beam search: At every iteration, only keep the k most promising search states and prune high-cost states

Shortcomings of PBMT

- Brittle design choices (Wu et al. 2016)
- Large phrase translation tables
- Inability to generalise (Kalchbrenner and Blunsom 2013)
  - i.e., similar phrase pairs do not share statistical weight in the models' estimation of their translation probabilities
  - Leading to general sparsity issues
- Imperfect translations

PBMT sample output

French	English (Google 2014)
Pour une journée au ski réussie à Samoëns, n'hésitez pas, consultez les prévisions météo sur la station, les conditions d'enneigement et l'ouverture des pistes. Profitez d'une bonne journée au ski à Samoens, si vous venez en week end ou pour une semaine, bénéficiez de prevision météo à 2 jours sur notre site et à plus long terme sur le site de Météo Chamonix !	For a successful day at Ski Samoens , please, check the forecast weather station on the snow conditions and opening tracks. Enjoy a good day skiing Samoens , if you come on a weekend weekend or for a week, get forecasting weather two days on our website and in the longer term on the site Weather Chamonix !

Part 3

NMT

Neural Machine Translation (NMT)

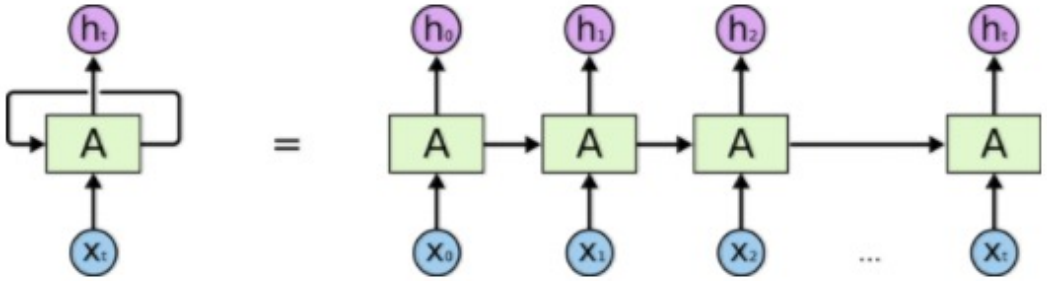
- Continuous representations (e.g., word2vec embeddings) for words and phrases are able to capture their morphological, syntactic and semantic similarity
- As in SMT, train on parallel corpora where sentences are aligned
- Maximise the probability of the sequence of tokens in the target language $y_1 \ldots y_m$ given the sequence of tokens in the source language $x_1 \ldots x_n$

$$P(y_1, \ldots, y_m | x_1, \ldots, x_n)$$

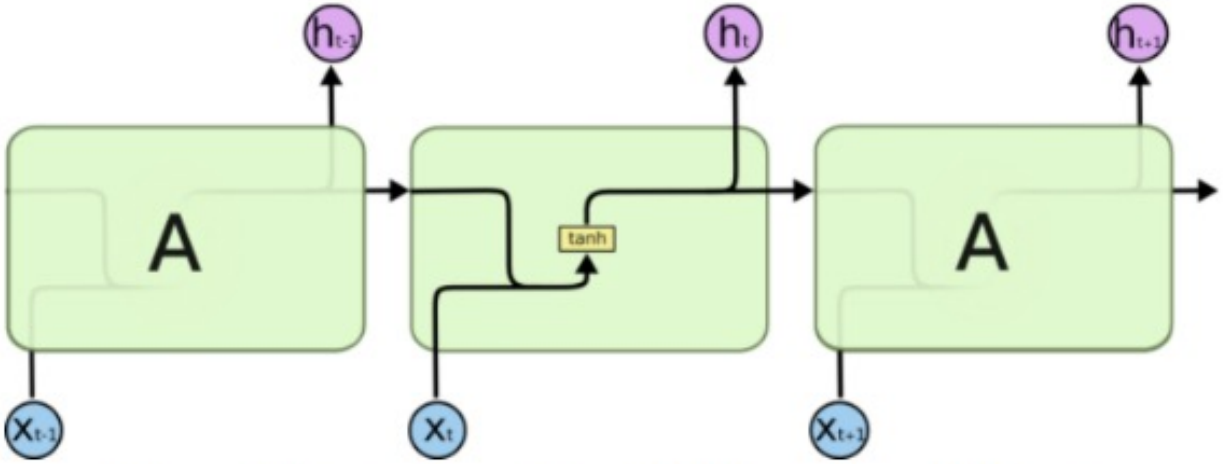
Basic architecture for NMT

- Encoder – decoder architecture
  - Aka sequence-to-sequence or seq2seq architecture
- 2 recurrent neural networks (RNNs) – one to consume the input text sequence and one to generate translated output text.

RNNs



An unrolled recurrent neural network.

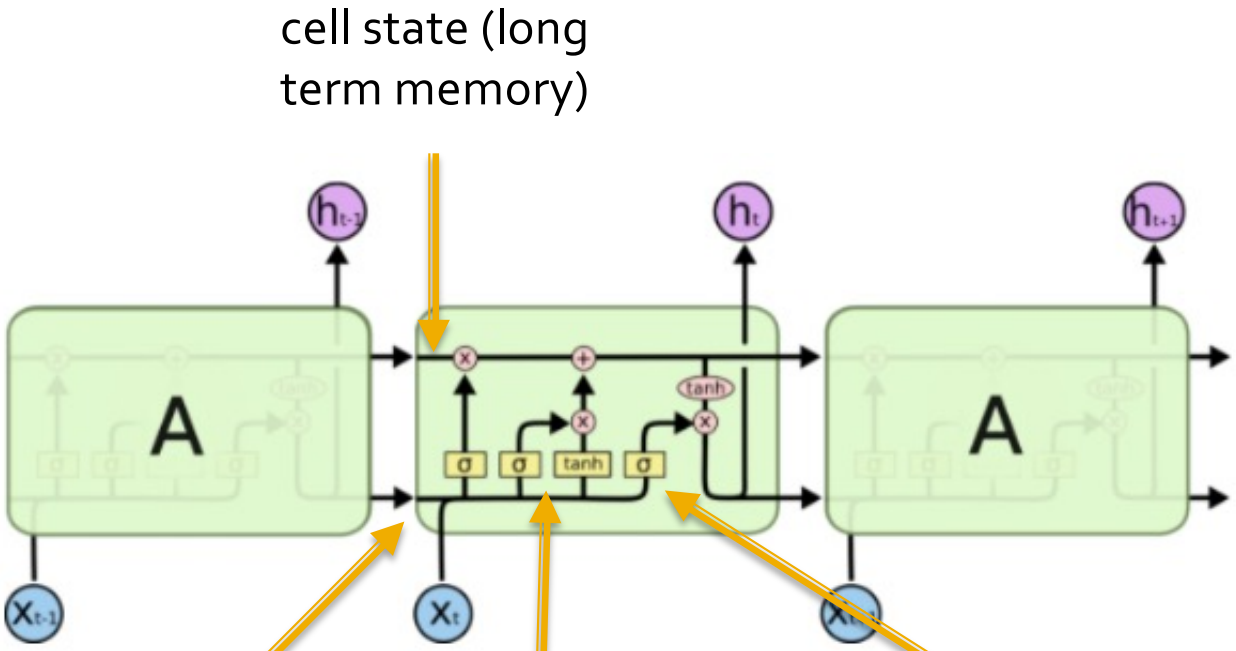


The repeating module in a standard RNN contains a single layer.

RNNs are very effective at learning language models i.e., P(E) the probability of a sentence in a given language. During training, the error (i.e., difference between output and next word) is back-propagated to update the weights used to combine X_t and h_{t-1} AND the representations of the words (X_t)

Long short term memory networks (LSTMs)

- Simple RNNs struggle with long term dependencies e.g., "He grew up in Spain. He speaks fluent …"



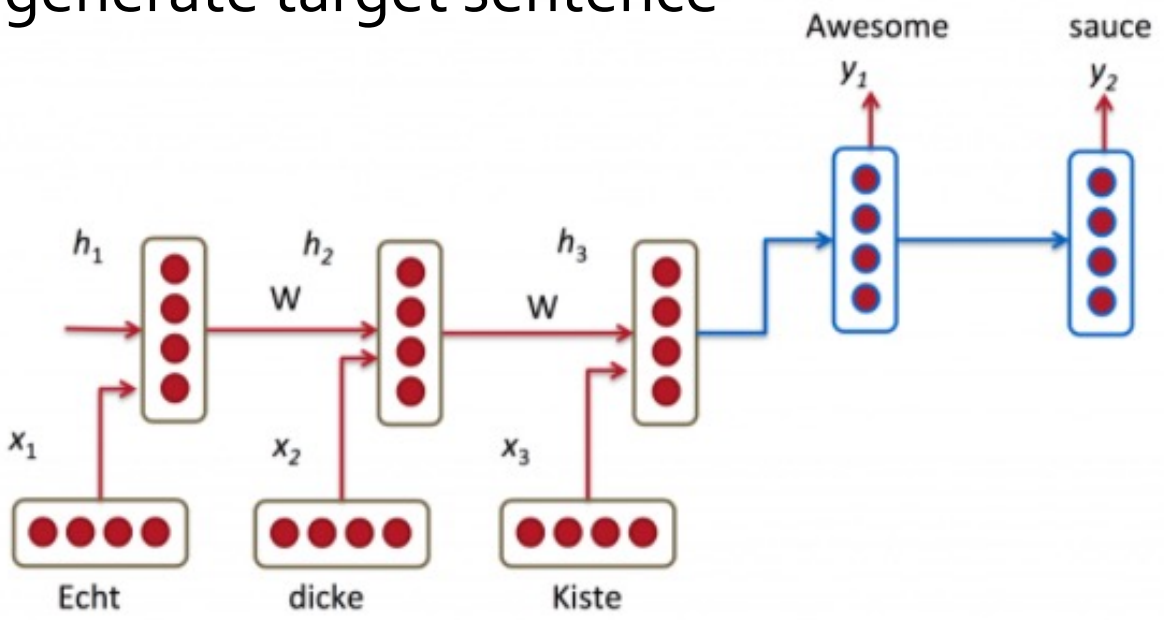
LSTMs overcome this problem by having 4 interacting layers in each repeated module.

Basic architecture for NMT

- RNN1, the encoder, builds a representation of the source sentence $x = x_1 \ldots x_n$

$$h = encoder(x)$$

- The output from RNN1 (after the complete source sentence has been read) is input to RNN2, the decoder to generate target sentence



$$y_{i+1} = decoder(h, y_1 \ldots y_i)$$

Encoder-decoder details

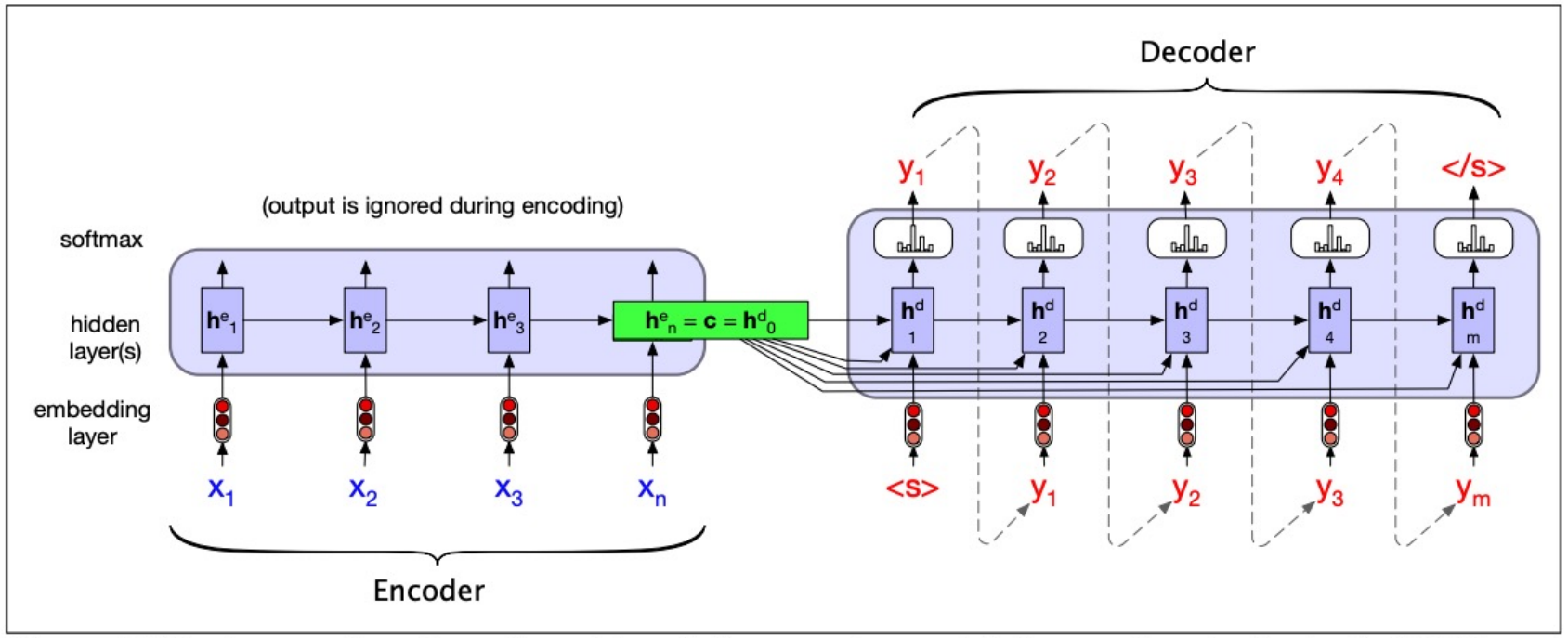


Figure 9.18 A more formal version of translating a sentence at inference time in the basic RNN-based encoder-decoder architecture. The final hidden state of the encoder RNN, h_n^e, serves as the context for the decoder in its role as h_0^d in the decoder RNN, and is also made available to each decoder hidden state.

Possible weaknesses

- Slow training and inference speed
- Ineffectiveness at dealing with rare words
- Output sentences that do not translate all words of the input sentence
- Difficulty in translating long sentences since the encoder output (or context) needs to encode the whole sentence
  - Information from start of sentence may be lost

Rare words (Luong et al. 2015)

- Due to computational constraints, NMT systems usually limited to top 30K-80K of most frequent words in each language
- Unknown/rare words can be translated using a dictionary or exact copy provided it is known which source word generated UNK token in target.
- Problem when sentence contains multiple rare words
- Luong et al. first use a word alignment of parallel corpora and annotate unknown words with positional information (e.g., UNK1)
- Output from NMT can then be post-processed

Subword tokenization

- Word vocabulary is huge and sparse
- Character vocabulary is small and dense, but lacking in semantic meaning
- Subword tokenization provides a compromise
- Frequent words tend to be a token whereas rare words will be broken down into subwords based on character n-grams
- Shared vocabulary for source and target languages – makes it easy to copy tokens like names from source to target
- Common algorithms include
  - BytePiece Encoding (BPE)
  - Wordpiece algorithm
  - Unigram / SentencePiece algorithm

Long sentences

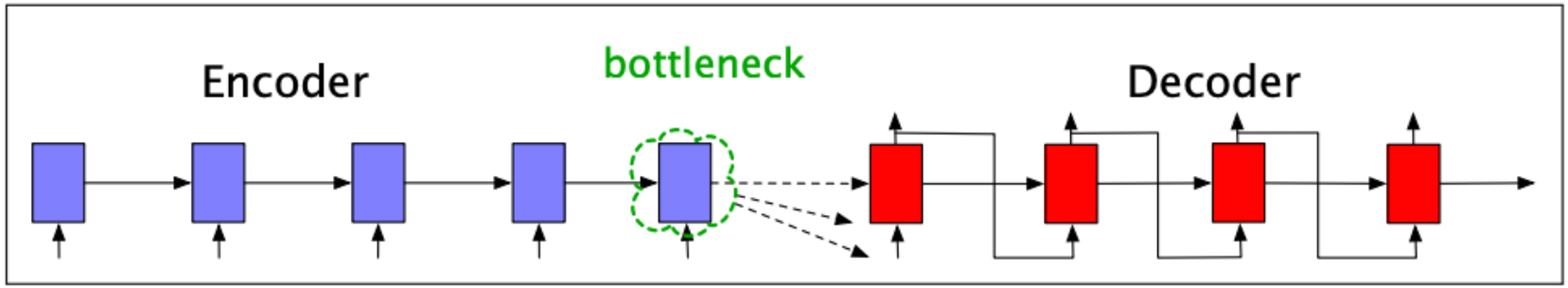


Figure 9.20 Requiring the context *c* to be only the encoder's final hidden state forces all the information from the entire source sentence to pass through this representational bottleneck.

- Attention (more on this next week) provides a way for the decoder to get information from all of the hidden states of the encoder rather than just the last hidden state

Attention

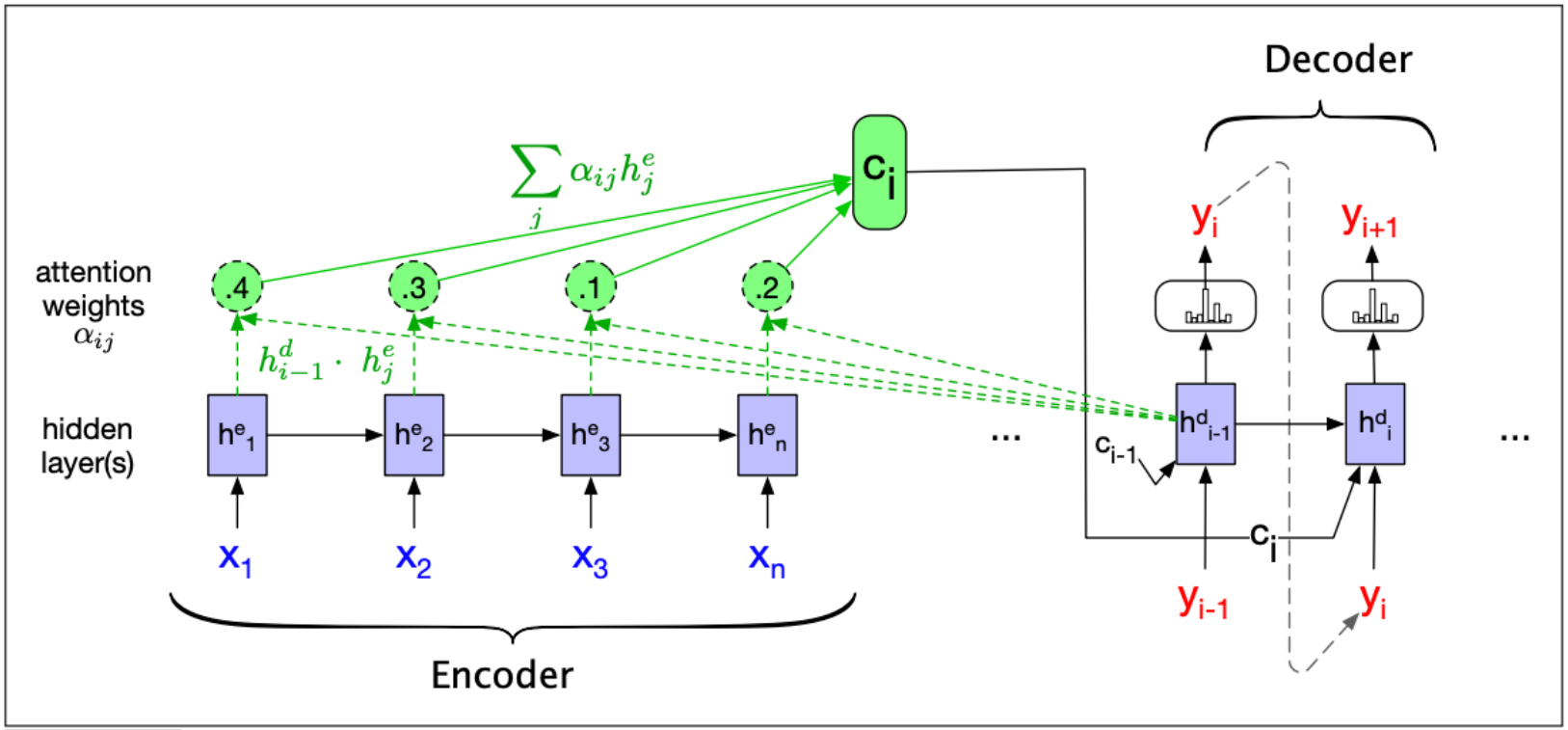
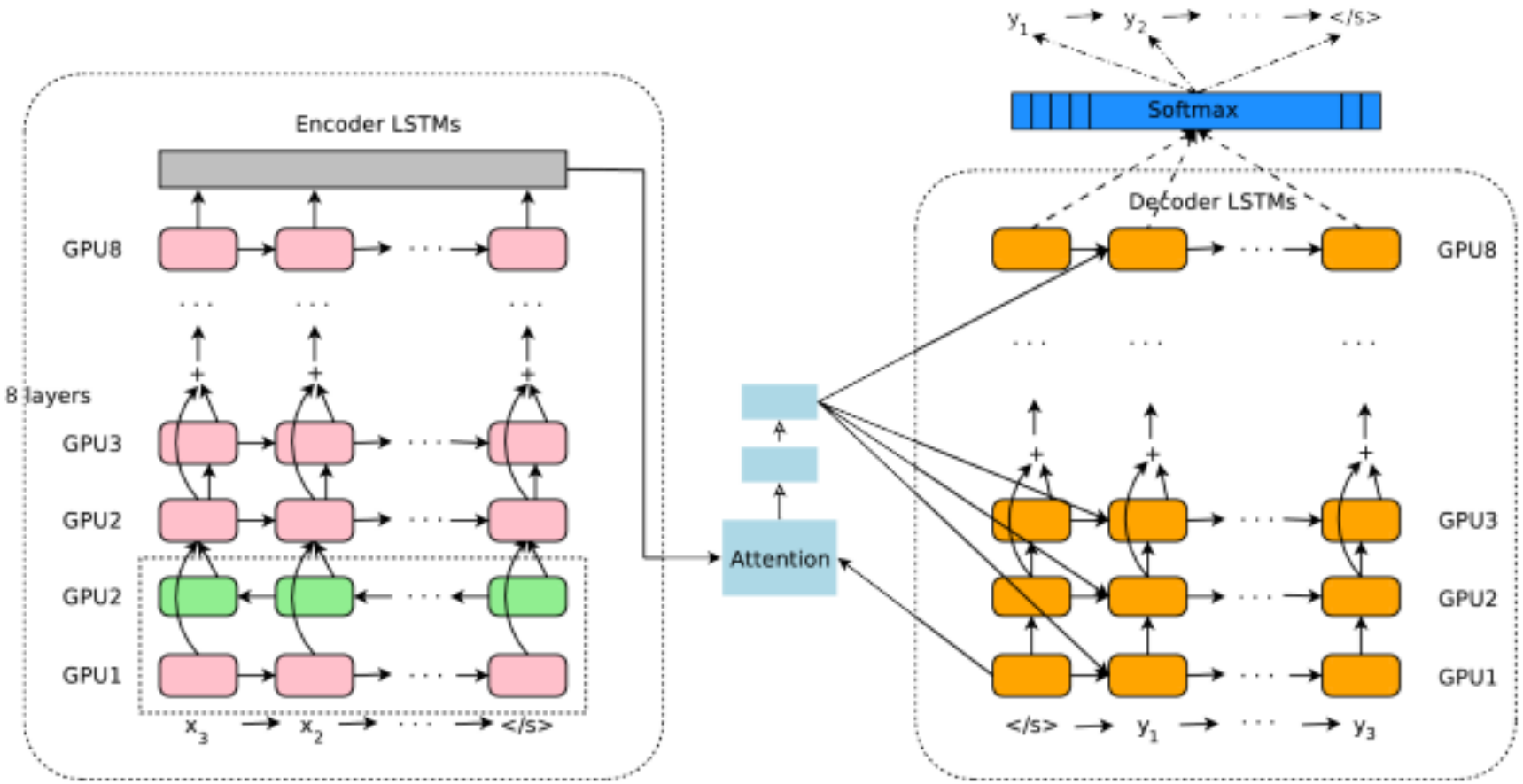


Figure 9.22 A sketch of the encoder-decoder network with attention, focusing on the computation of $\mathbf{c}_i$. The context value $\mathbf{c}_i$ is one of the inputs to the computation of $\mathbf{h}_i^d$. It is computed by taking the weighted sum of all the encoder hidden states, each weighted by their dot product with the prior decoder hidden state $\mathbf{h}_{i-1}^d$.

Google NMT (GNMT)

- Recurrent networks are LSTMs with attention (8 layers - residual connections between layers to encourage gradient flow)
- For parallelism, the attention from the decoder network connect to top layer of encoder network
- Low-precision arithmetic for inference, accelerated using special hardware (Google's TPU)
- Rare words dealt with using sub-word units (balancing flexibility of single characters with efficiency of full words)
- Beam search techniques includes a length normalization procedure and a coverage penalty to encourage model to translate all of the input

GNMT Architecture



Transformers and LLMs in MT?

- Transformers generally have higher performance than LSTMS and GRUs
- Generally replacing seq2seq architectures
- More on this in weeks 8-10

- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9355969
- https://arxiv.org/abs/2209.07417

Open questions

- High resource vs low resource languages

References

- Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models, In *EMNLP*
- Philip Koehn. 2004. Pharoah: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *AMTA*
- Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals and Wojciech Zaremba. 2015. Addressing the Rare Word Problem in Neural Machine Translation. In *ACL*
- Ilya Sutskever, Oriol Vinyals and Quoc Le. Sequence to Sequence Learning with Neural Networks. In *NIPS*
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le and Mohammad Norouzi. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv Oct 2016