CS 546 – Advanced Topics in NLP

Dilek Hakkani-Tür

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

CONVAI

CDS

Siebel School of Computing and Data Science

Topics for Today

- Encoder-Decoder Architecture
- Beam Search
- Case Study: Response generation in E2E Dialogue Systems
- Limitations of S2S models & Solutions
- Attention
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

Readings

- Ch 11 of the Dive into Deep Learning book

Variable Length Sequences

- Inputs can have variable length!

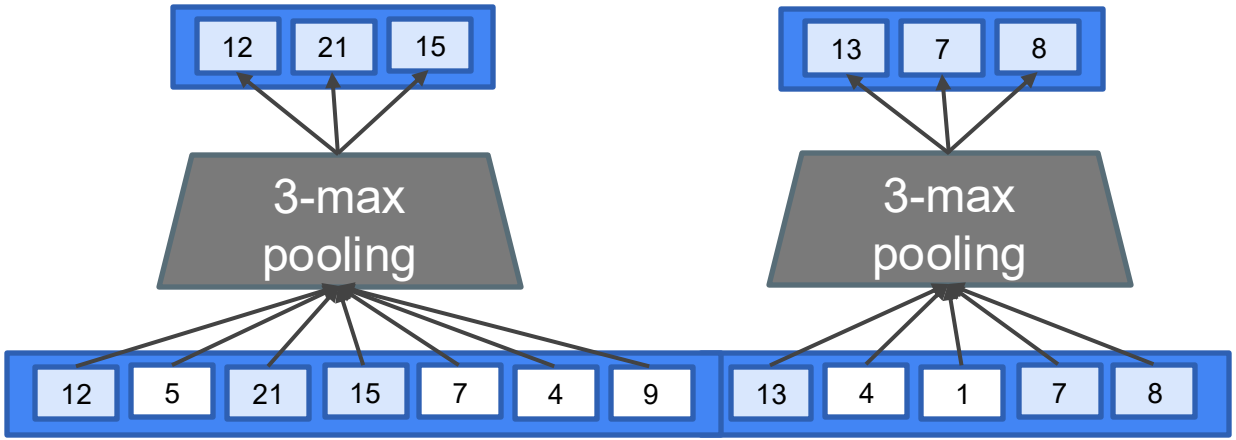
- V = {the, a, chased, cat, dog, is, cute}

		Bag of words:
- the cat is cute	→	[1 0 0 1 0 1 1]
- the cat chased a dog	→	[1 1 1 1 1 0 0]
- the dog chased a cat	→	[1 1 1 1 1 0 0]

Bag of words approach represents variable length inputs with vectors of size |V|. Works well for tasks where word ordering is not important.

Variable Length Sequences (cont.)

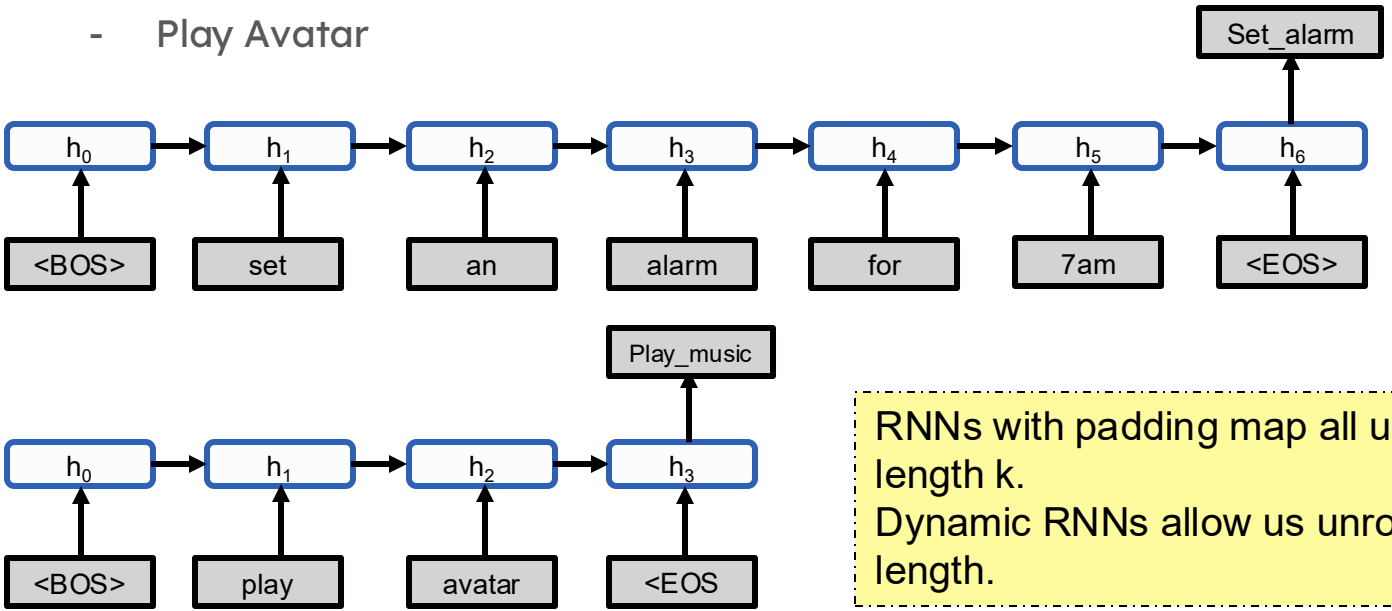
- CNNs with k-max pooling:



k-max pooling maps all inputs to a fixed number of vectors.

Variable Length Sequences (cont.)

- RNNs for classification, i.e., intent:
  - Set an alarm for 7am
  - Play Avatar



RNNs with padding map all utterances to length k.
Dynamic RNNs allow us unroll to utterance length.

Variable Length Sequences (cont.)

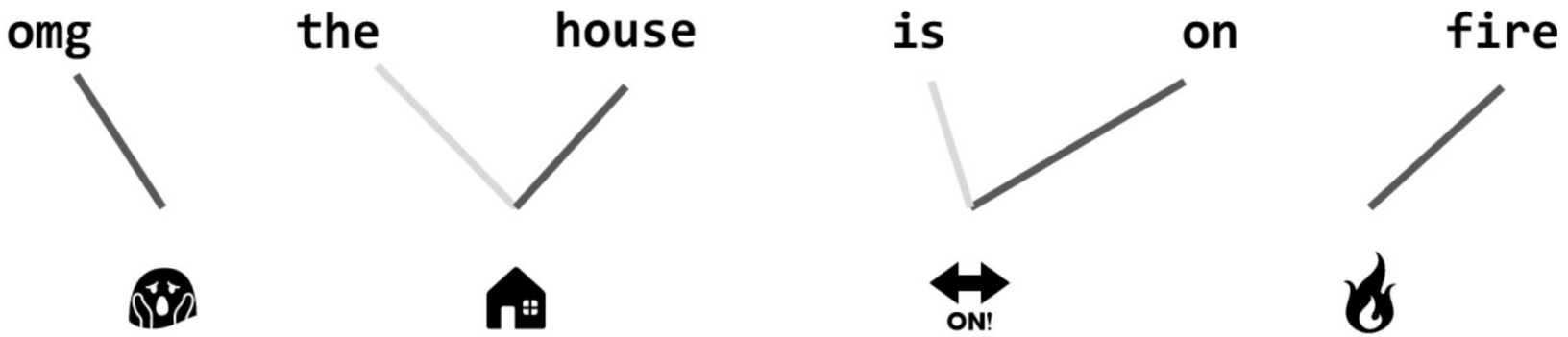
- What if the outputs are a list of tokens of variable length?

- Examples:

Task	Input	Output
nasılsın	Nasılsın? (in Turkish)	How are you?
Summarization	1 or more documents	Summary
Dialogue Systems	Previous conversation A: How are you? B: Thanks, I am doing ok. And you?	Next response A: Thanks, I am fine. Have you seen the Parasite movie? Wanna go see it together tomorrow?

Example: Emoji Translation

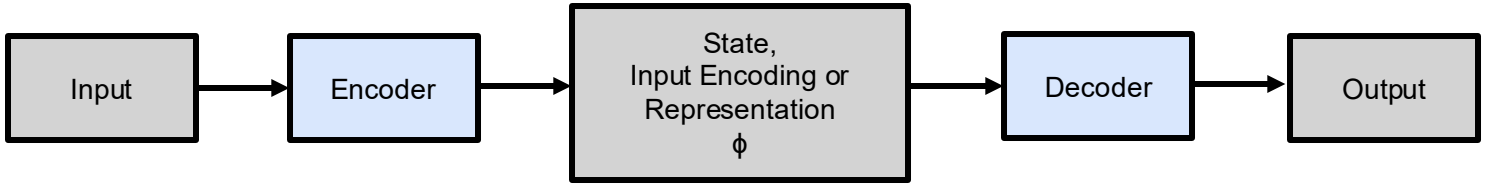
- Aiming to converting messages typed on a phone keyboard to emojis.
- Alignment between the input and the output:



Other examples: machine translation, question answering, dialogue response generation, …

Encoder-Decoder Architecture

- Composition of two models:



- Encoder aims to capture important information for the task
- ϕ is usually a vector
- Decoder aims to decode the encoded input to target output
- Trained jointly

Earlier Work: Interlingua-based Machine Translation

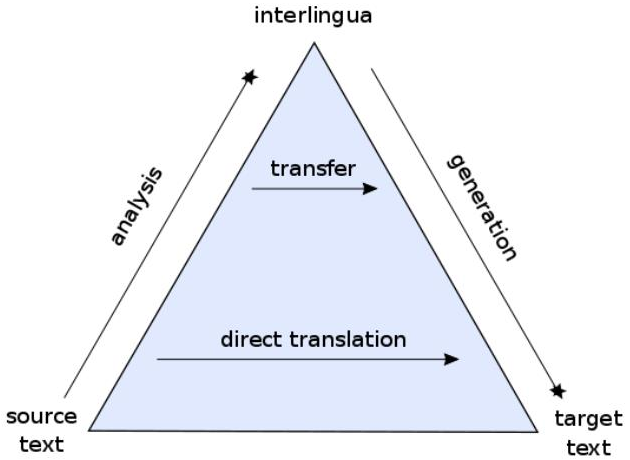


Image from Wikipedia: https://en.wikipedia.org/wiki/Transfer-based_machine_translation

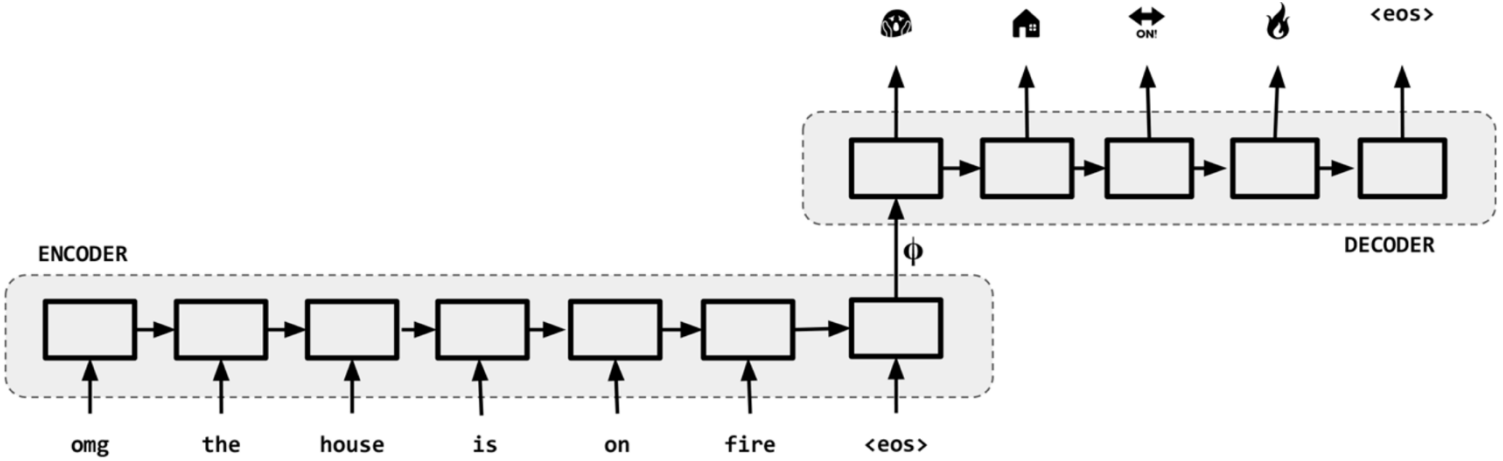
Interlingua:

- Language independent meaning representation
- makes explicit the distinctions necessary for successful translation

Encoder-decoder models aim to learn representations from data: “Representation learning”

S2S Learning with Neural Networks

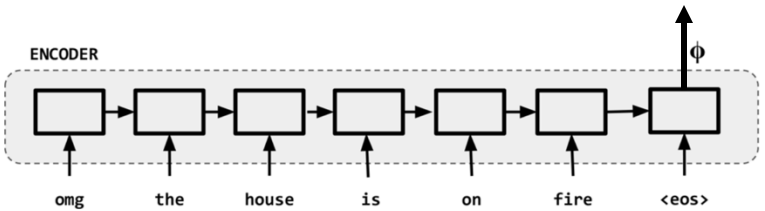
- For NL sequences, the encoder and decoder could both be RNNs.
- Sequence-to-sequence (S2S) models
- (Sutskever et al, 2014)



S2S Learning with Neural Networks: The Encoder

- Input: $x_1,\ldots,x_T$ where x_t is the t^{th} word
- The transformation of the RNN's hidden states, denoted by f: $\mathbf{h}_t=f(\mathbf{x}_t,\mathbf{h}_{t-1})$
- The encoder captures information and encodes it into the context vector ϕ with a function q:$\phi=q(\mathbf{h}_1,\ldots,\mathbf{h}_T)$

$$q(\mathbf{h}_1,\ldots,\mathbf{h}_T)=\mathbf{h}_T$$



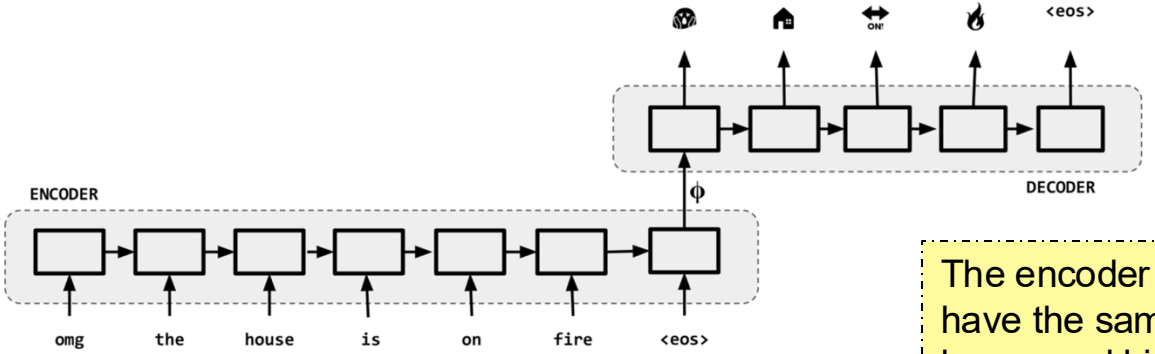
S2S Learning with Neural Networks: The Decoder

- The context vector, ϕ, encodes the information from the complete input sequence $x_1,\ldots,x_T$.
- If given outputs in the training set are $y_1,\ldots,y_{T'}$, at each timestep t',

$$P(y_{t'}|y_1,\ldots,y_{t'-1}, \phi)$$

- Another RNN as decoder, with hidden states s_t

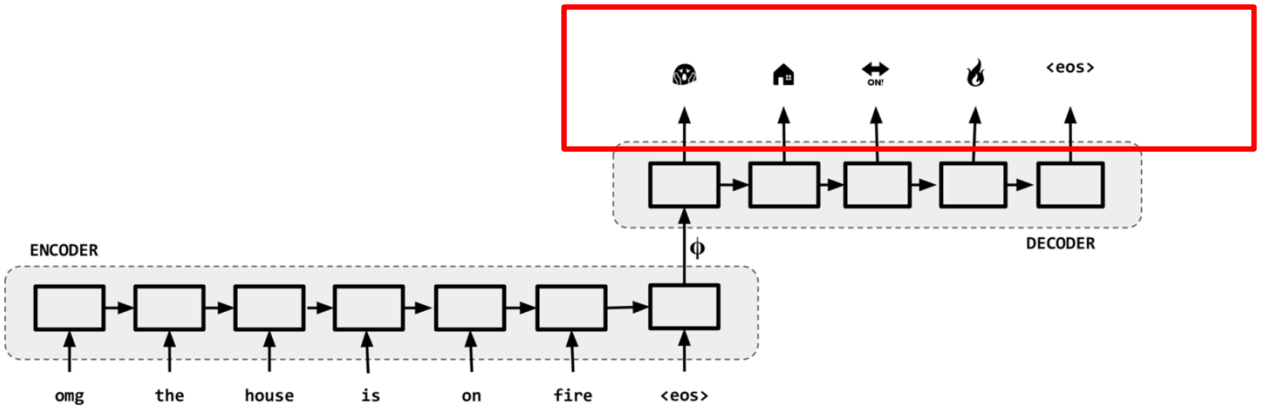
$$\mathbf{s}_{t'}=g(\mathbf{y}_{t'-1}, \phi,\mathbf{s}_{t'-1})$$



The encoder and decoder RNNs have the same numbers of layers and hidden unit size.

S2S Learning with Neural Networks (cont.)

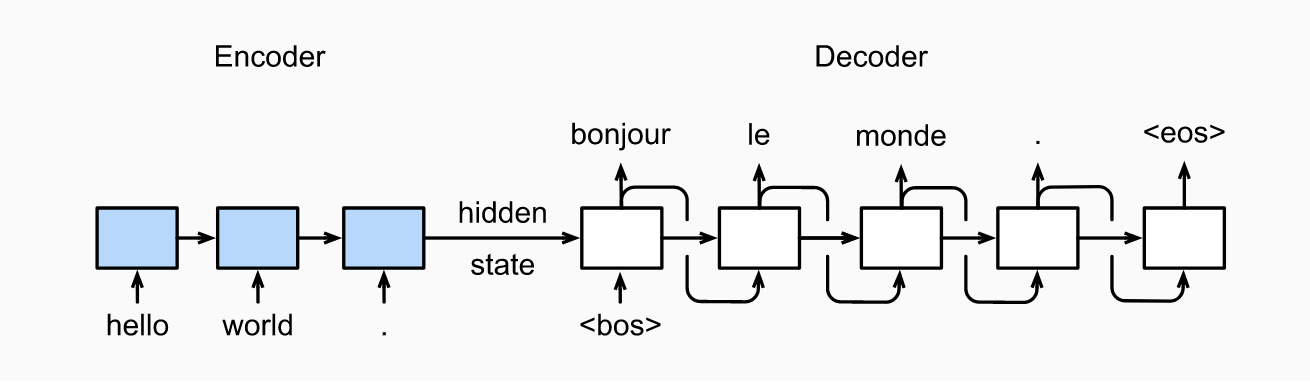
- Decoder has a dense layer after the RNN layers, where the hidden size is the vocabulary size.
- The dense layer predicts the confidence score for each word.
- We can use softmax on the output scores
- And use cross entropy as the loss.



S2S Learning with Neural Networks (cont.)

- Simplest method for predictions, greedy search.

$$y_{t'}=\text{argmax}_{y\in Y}P(y|y_1,\dots,y_{t'-1},\phi)$$



Generate until the “<eos>” symbol is detected, or the output sequence has reached the max length T.

Issue with Greedy Search

- The conditional probability for generating an output sequence

$$\prod_{t'=1}^{T'} P(y_{t'} | y_1, \ldots, y_{t'-1}, \phi)$$

Timestep	1	2	3	4
A	0.5	0.1	0.2	0.0
B	0.2	0.4	0.2	0.2
C	0.2	0.3	0.4	0.2
<eos>	0.1	0.2	0.2	0.6

P("ABC<eos>"|…) = 0.048

Timestep	1	2	3	4
A	0.5	0.1	0.1	0.1
B	0.2	0.4	0.6	0.2
C	0.2	0.3	0.2	0.1
<eos>	0.1	0.2	0.1	0.6

P("ACB<eos>"|…) = 0.054

No guarantee that optimal sequence will be obtained with greedy search!

Exhaustive Search

- Examination of all possible sequences.
- Computational overhead: $O(|Y|^{T'})$
- For example, if $|Y|$=10000 and T'=10, we will need to evaluate $10000^{10} = 10^{40}$ sequences

Exhaustive search is too expensive even with small vocabulary sizes!

Topics for Today

- Encoder-Decoder Architecture
- Beam Search
- Case Study: Response generation in E2E Dialogue Systems
- Limitations of S2S models & Solutions
- Attention
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

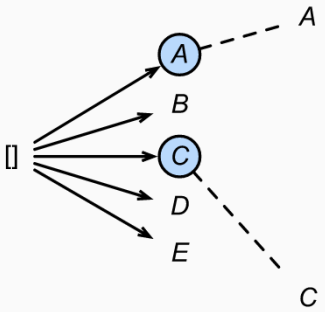
Beam Search

- Beam-size = k
- At time step 1, select top k candidates
- For each subsequent timestep:
  - Determine the $k|Y|$ possible output sequences based on the k candidate output sequences from the previous timestep
  - select the top k output sequences with the highest conditional probability amongst those.

Beam Search (cont.)

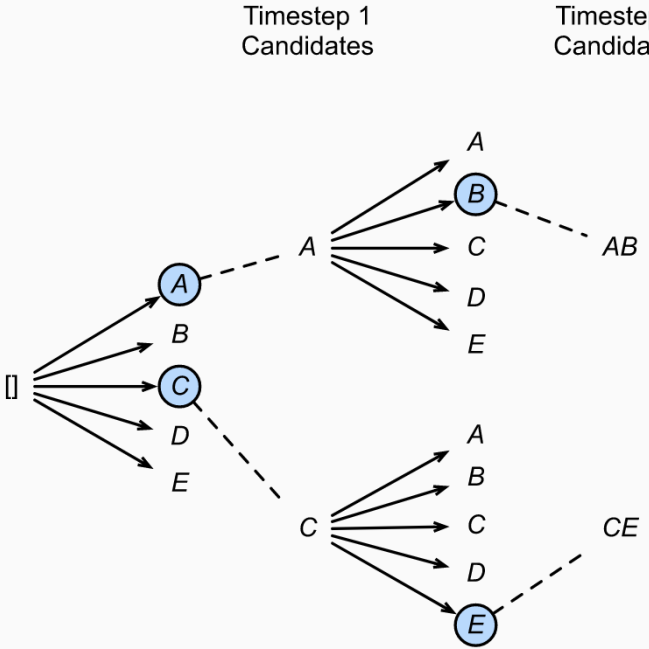
k=2

Timestep 1
Candidates

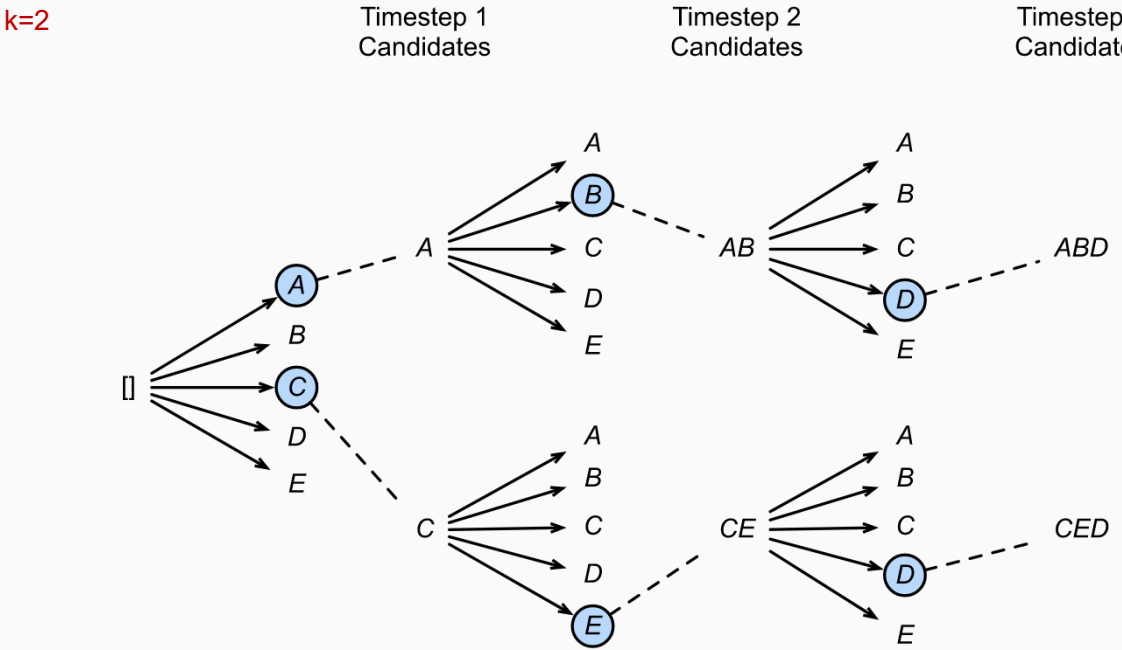


Beam Search (cont.)

k=2



Beam Search (cont.)



Beam Search (cont.)

- The computational overhead of beam search is between greedy search and exhaustive search.

- Greedy search can be treated as a beam search with a beam size of 1.

- Beam search strikes a balance between computational overhead and search quality using a flexible beam size of k.

Topics for Today

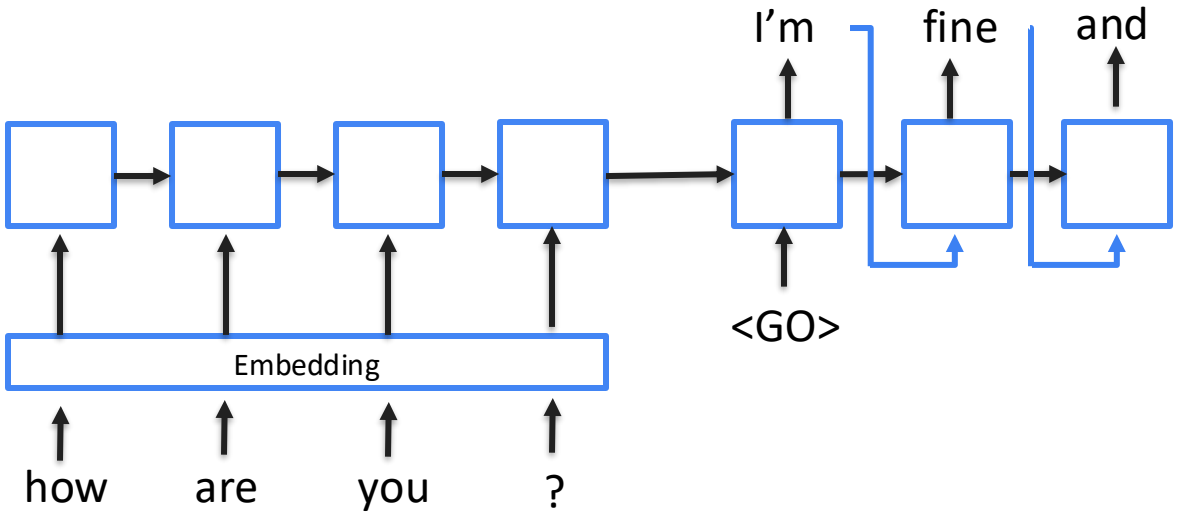
- Encoder-Decoder Architecture
- Beam Search
- Case Study: Response generation in E2E Dialogue Systems
- Limitations of S2S models & Solutions
- Attention
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

Response Generation in E2E Dialogue Systems

- **Aim:** Learn to generate a response from large conversational datasets.
- **Inference time:** Given the conversation context, generate a natural and appropriate response.
- Modeled with E2E architectures.

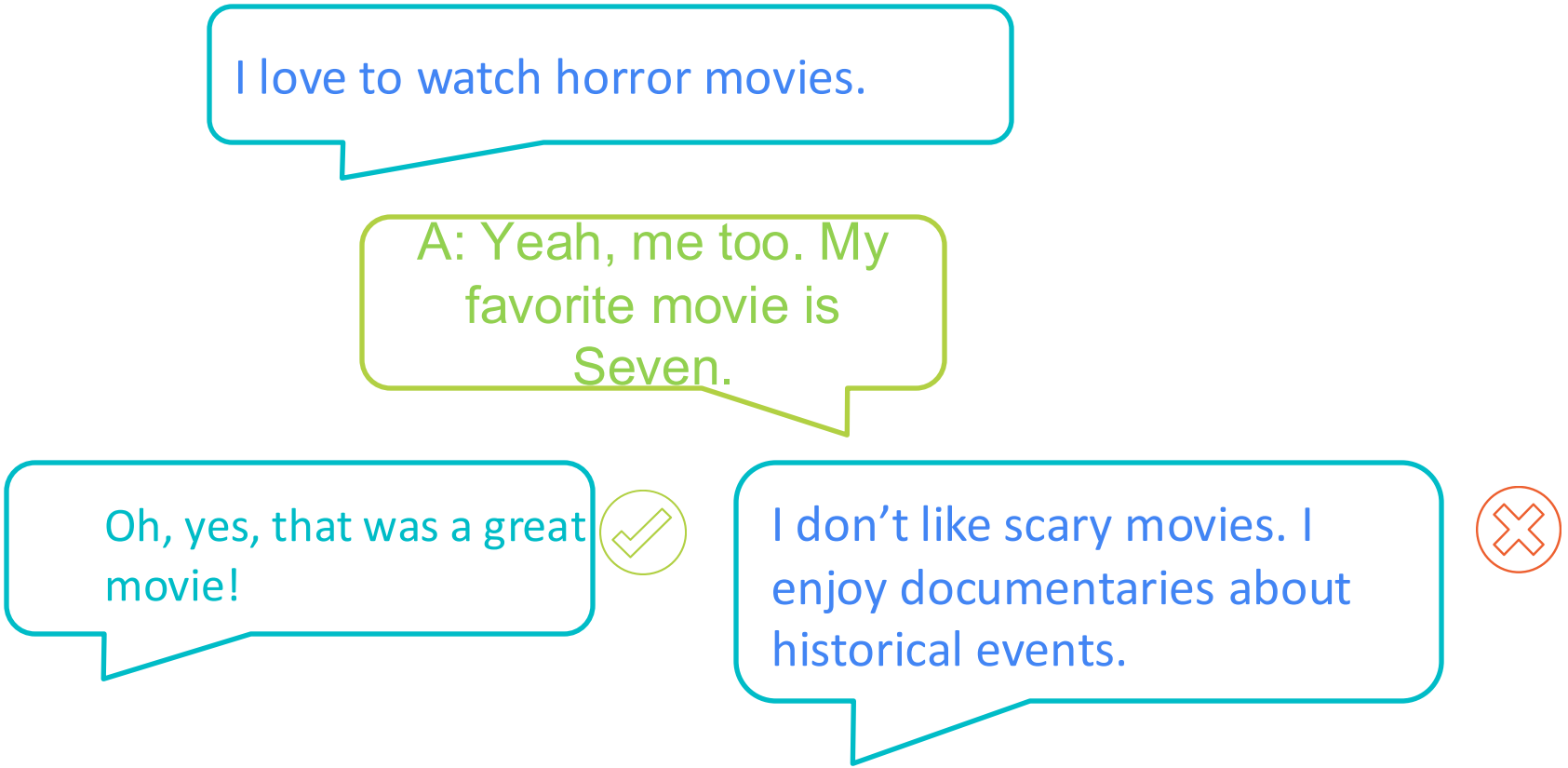
Sequence-to-Sequence Models for RG

Modeling conversations as request-response pairs.



Why is context important for RG?

Why is context important for RG?



Representations of Conversation Context

The complete conversation context can be treated as a long sequence.

U: Hi, how have you been?
S: I am doing great, and you?
U: Good, thanks, I am looking for an interesting movie to see.
S: Deepwater Horizon is showing at Angelika.
U: What else is there?
S: Not sure. Did you know more than 200 million gallons of oil spilled into the Gulf of Mexico during that time?
U: Oh, wow, I didn't know that, I cannot even imagine how much that would be. What are the showtimes?
S: You can see it today at 3:40pm and 5:45pm.

U: "Let's do 3:40"

Representations of Conversation Context

The complete conversation context can be treated as a long sequence.

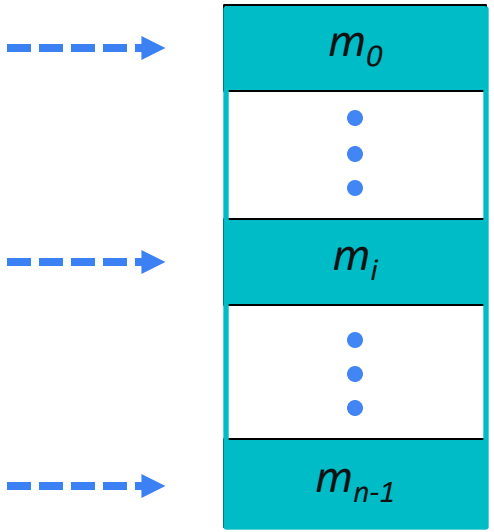
It could also be limited length window, defined by # of words / characters.

U: Hi, how have you been?
S: I am doing great, and you?
U: Good, thanks, I am looking for an interesting movie to see.
S: Deepwater Horizon is showing at Angelika.
U: What else is there?
S: Not sure. Did you know more than 200 million gallons of oil spilled into the Gulf of Mexico during that time?
U: Oh, wow, I didn't know that, I cannot even imagine how much that would be. What are the showtimes?
S: You can see it today at 3:40pm and 5:45pm.

U: "Let's do 3:40"

Representations of Conversation Context

U: "i d like to purchase tickets to see deepwater horizon"
S: "for which theatre"
U: "angelika"
S: "you want them for angelika theatre?"
U: "yes angelika"
S: "how many tickets would you like ?"
U: "3 tickets for saturday"
S: "What time would you like ?"
U: "Any time on saturday is fine"
S: "okay , there is 4:10 pm , 5:40 pm and 9:20 pm"
U: "Let's do 5:40"



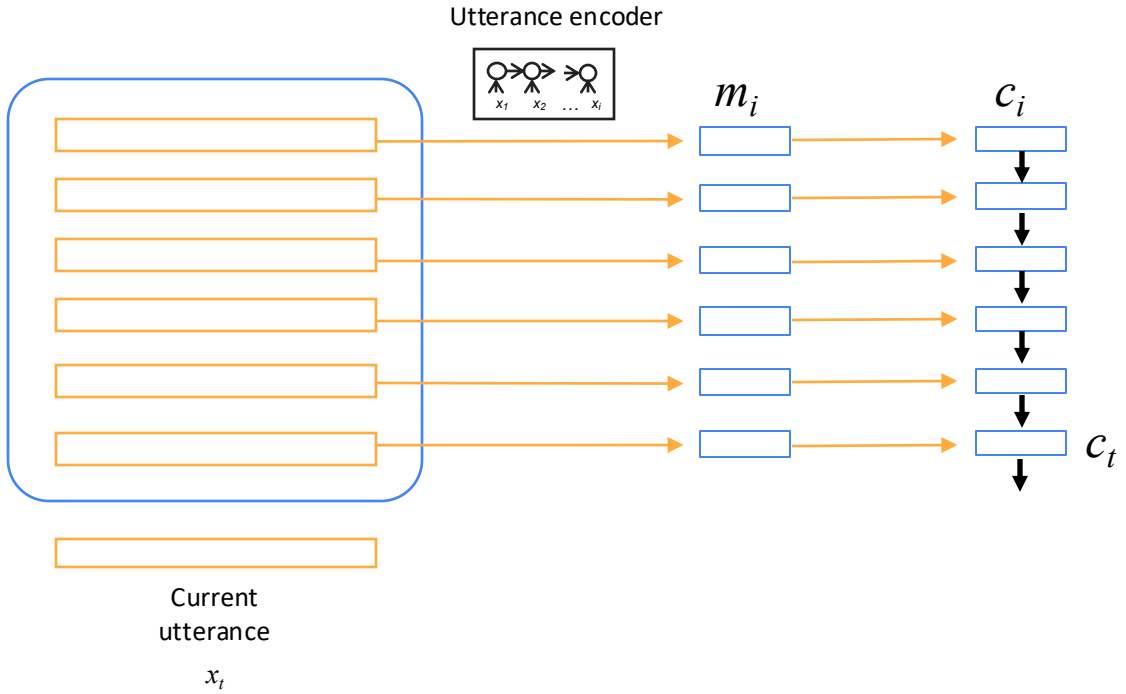
Representations of Conversation Context – Hierarchical RNNs

1. Sentence Encoding

$$m_t = RNN_1(x_t)$$

2. Context Encoding

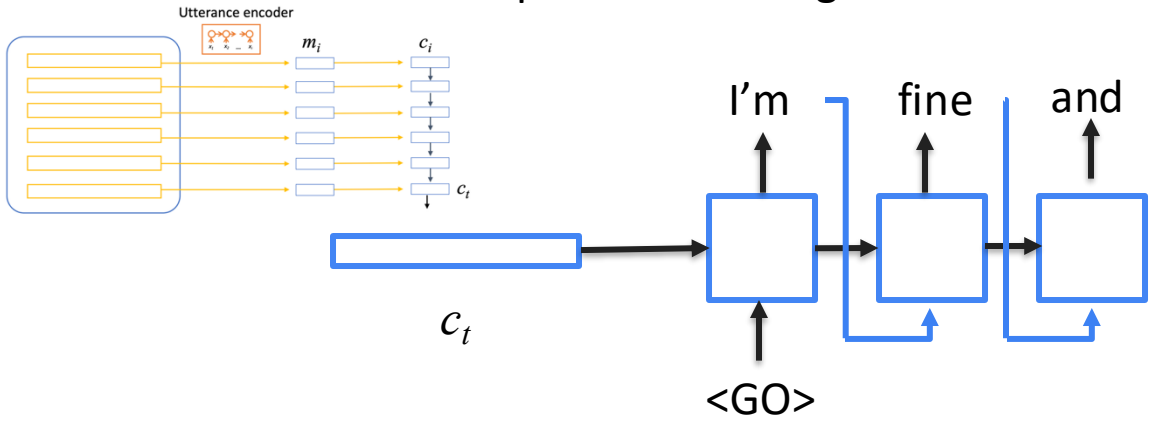
$$c_t = f(c_{t-1}, m_{t-1})$$



Serban et al, AAAI 2016. Building end-to-end dialogue systems using generative hierarchical neural network models.

Representations of Conversation Context – Hierarchical RNNs

- The hierarchical RNN encodes the complete conversation context.
- The context vector is the input for decoding.



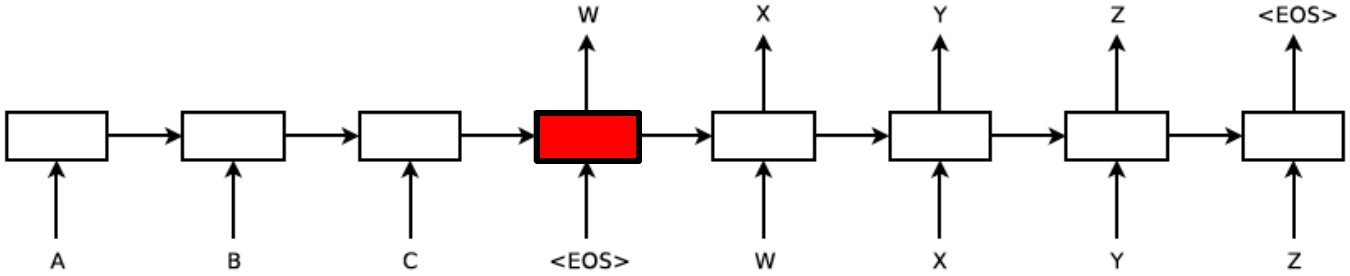
Serban et al, AAAI 2016. Building end-to-end dialogue systems using generative hierarchical neural network models.

Topics for Today

- Encoder-Decoder Architecture
- Beam Search
- Case Study: Response generation in E2E Dialogue Systems
- Limitations of S2S models & Solutions
- Attention
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

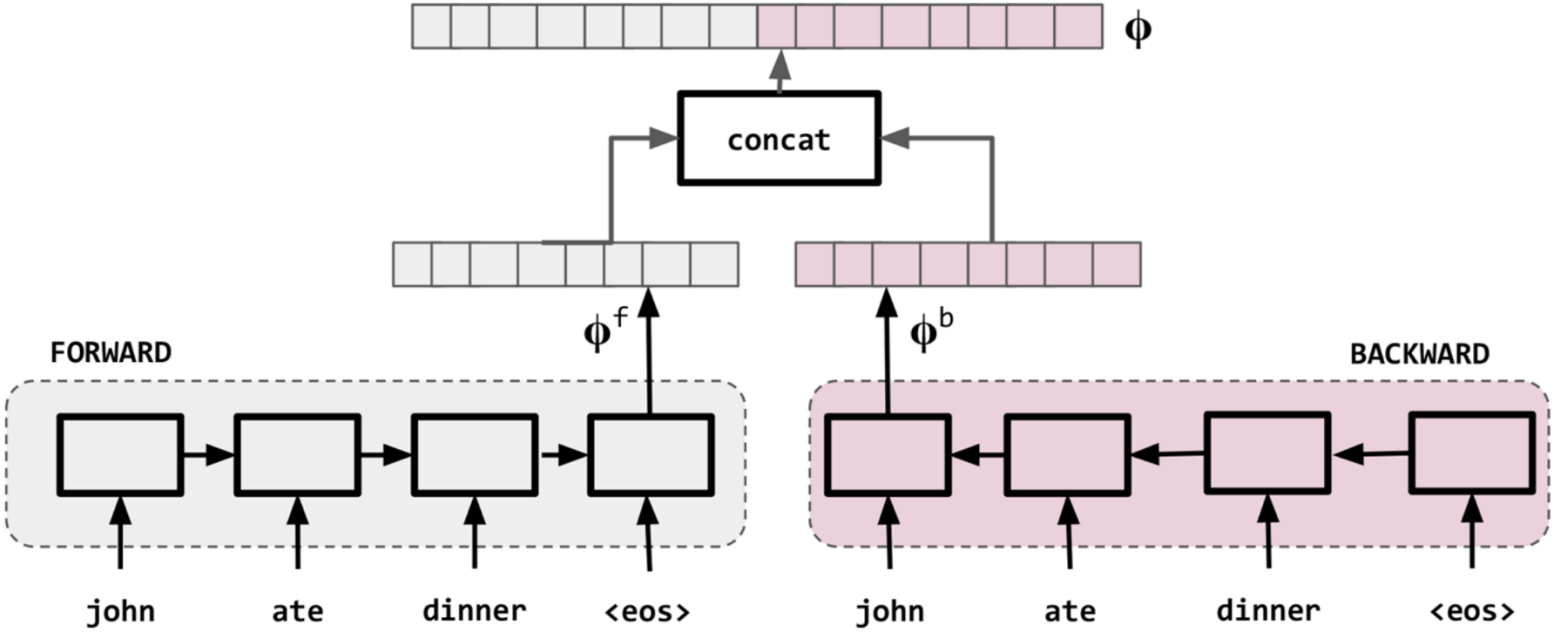
S2S Models: Limitations of Encoding

- The final embedding of the input should capture all the information!



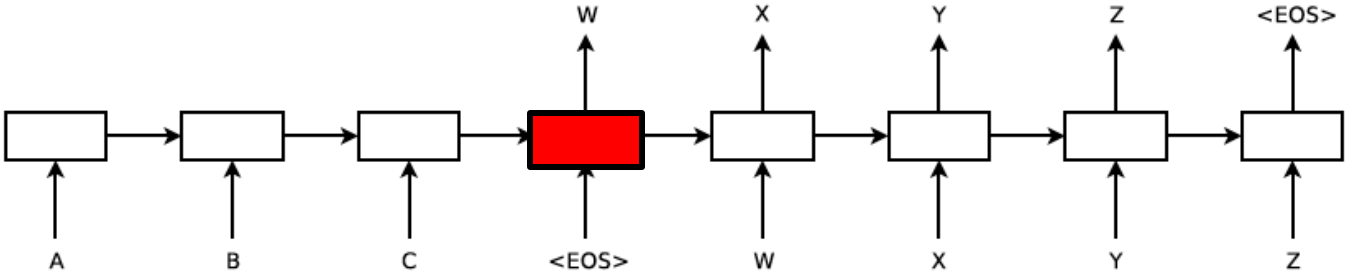
- Encoding of long sequences could be problematic.
  - Even with LSTM/GRU models

Bidirectional Encoder



S2S Models: Limitations of Encoding

- The final embedding of the input should capture all the information!



- Encoding of long sequences could be problematic.
  - Even with LSTM/GRU models
- **Hierarchical models can also be helpful.**
- **Some tasks (i.e., machine translation) require “attending” to specific words.**

Issue with S2S Models

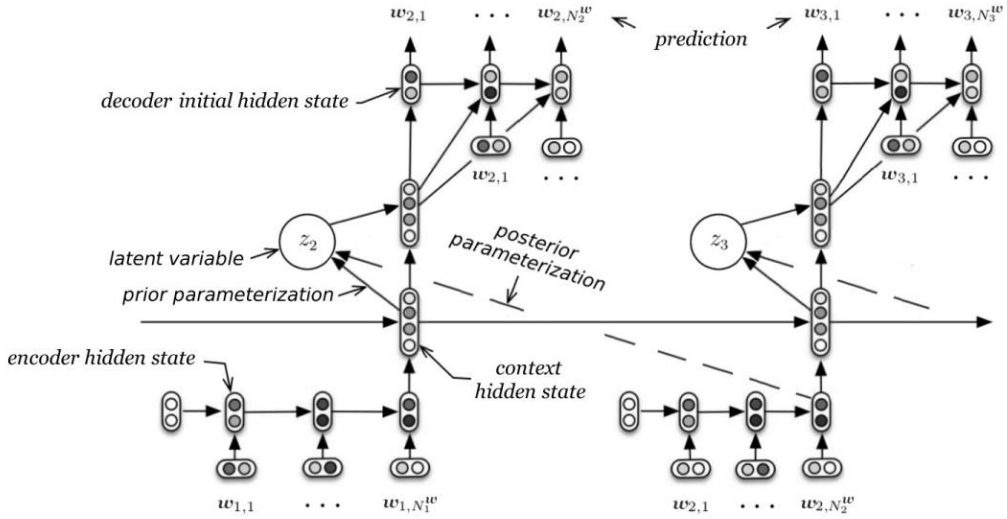
- Maximization based decoding methods, such as beam search, result in generic and bland utterances:
  “Absolutely!” “That is interesting!” “Ok”

- Often times with repetition
  “I I I am fine. I am fine.”

- How can we improve the diversity of the responses?

Latent Variable Encoder-Decoder Model

- Encoder: hierarchical recurrent neural networks to represent conversation context.
- Injecting stochastic noise during decoding for response generation.
- A latent variable at the decoder, allows to model hierarchically-structured sequences in a two-step generation process:
  - sampling the latent variable
  - generating the output sequence—while maintaining long-term context.



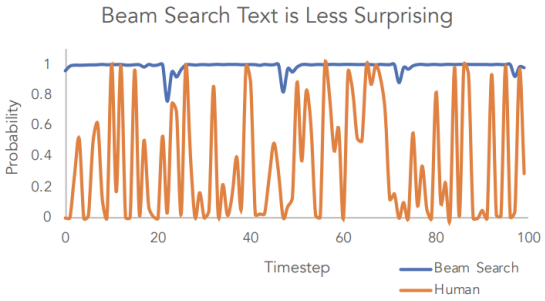
Serban et al, 2016, A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues.

Sampling Methods During Decoding

- To create diversity, sample, rather than picking the most probable token.

- Top- k sampling:
  - Keep only the top k tokens, re-normalize, then sample.
  - Prevents picking extremely unlikely tokens

Sampling Methods During Decoding (cont.)

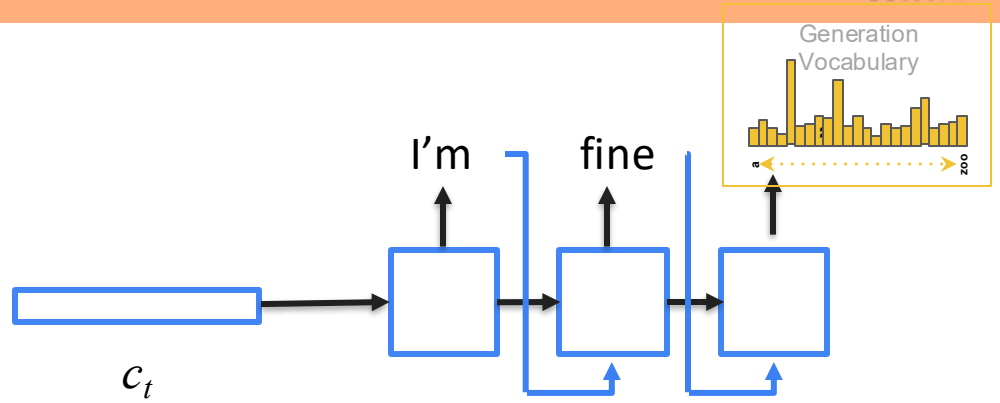


Beam Search

...to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and to provide an overview of the current state-of-the-art in the field of computer vision and machine learning, and...

Human

...which grant increased life span and three years warranty. The Antec HCG series consists of five models with capacities spanning from 400W to 900W. Here we should note that we have already tested the HCG-620 in a previous review and were quite satisfied With its performance. In today's review we will rigorously test the Antec HCG-520, which as its model number implies, has 520W capacity and contrary to Antec's strong beliefs in multi-rail PSUs is equipped...



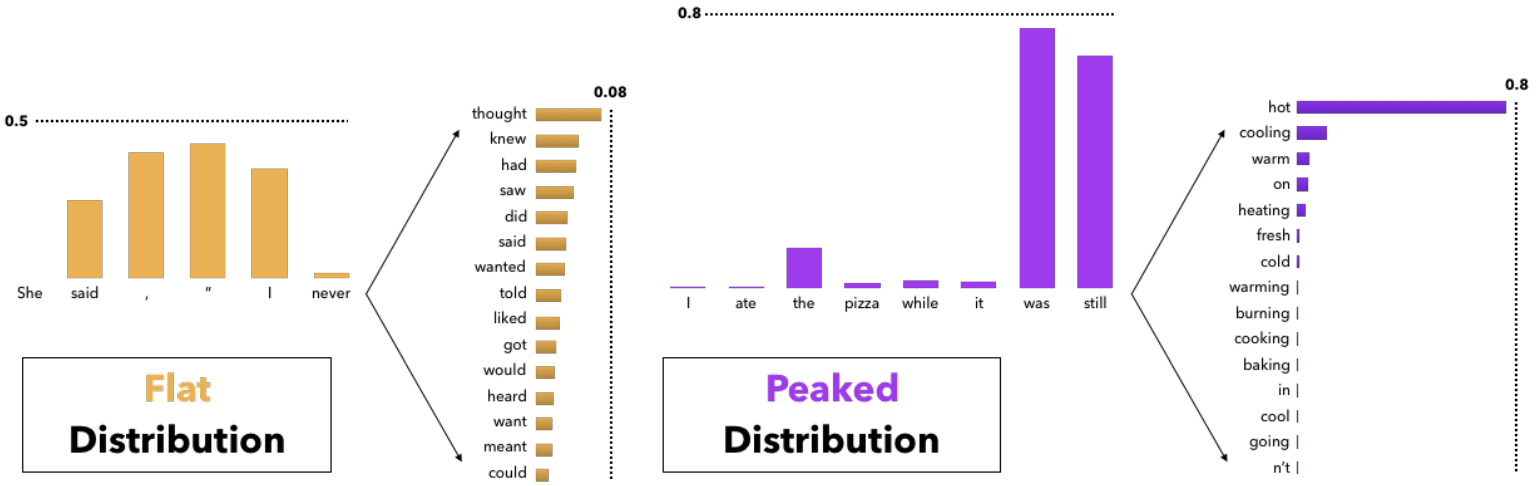
- Given the same context, humans generate words of varying probability (according to model), whereas beam search selects the top words.

- Nucleus Sampling

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p.$$

Holtzmann et al, ICLR 2020. The curious case of neural text degeneration.

Sampling Methods During Decoding (cont.)



In **Nucleus Sampling**, the number of candidates considered rises and falls dynamically. In **Top-** k **Sampling**, the number of candidates is fixed to k, which could be sub-optimal across varying contexts. If k is small, generated text will be bland, if k is large, then the list will include inappropriate candidates.

Holtzmann et al, ICLR 2020. The curious case of neural text degeneration.

Sampling Methods During Decoding (cont.)

- Another common approach, shape the probability distribution through temperature.
- Given the logits, $u_{l:|V|}$ and temperature t, the softmax is re-estimated as:

$$p(x = V_l|x_{1:i-1}) = \frac{\exp(u_l/t)}{\sum_{l'} \exp(u'_l/t)}.$$

$t < 1$: Sharper distribution → more deterministic
$t > 1$: Flatter distribution → more random

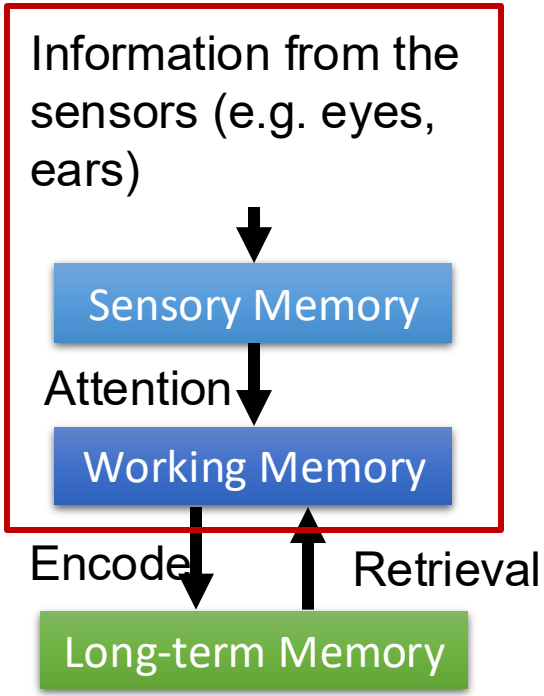
Setting temperature, t, in [0,1) skews the probability distribution towards high probability events. Shown to improve quality at the cost of reducing diversity.

Holtzmann et al, ICLR 2020. The curious case of neural text degeneration.

Topics for Today

- Encoder-Decoder Architecture
- Beam Search
- Case Study: Response generation in E2E Dialogue Systems
- Limitations of S2S models & Solutions
- Attention
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

Attention and Memory: Types of Memory

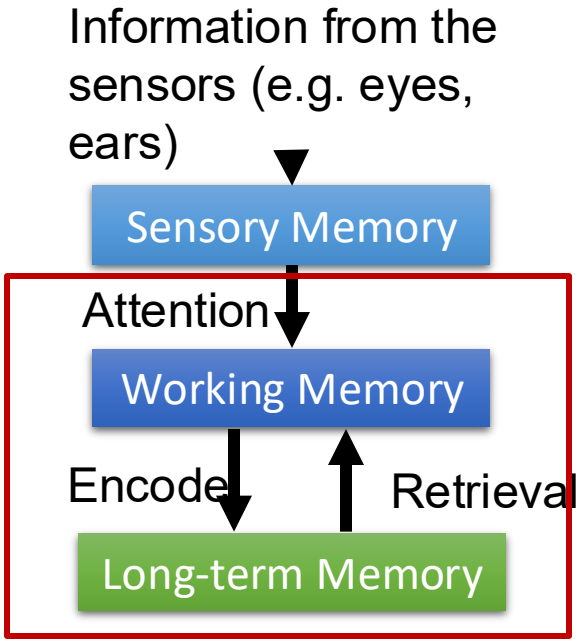


When the input is a very long sequence or an image

→ Pay attention on part of the input object each time

https://en.wikipedia.org/wiki/Memory

Attention and Memory: Types of Memory



When the input is a very long sequence or an image

→ Pay attention on part of the input object each time

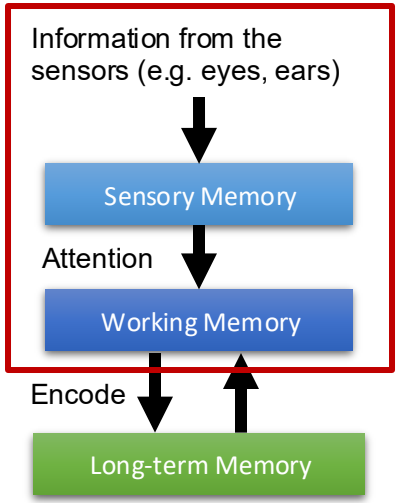
When access to longer term memory is needed

→ We can retrieve contents and use attention mechanisms

https://en.wikipedia.org/wiki/Memory

Attention over Context in Machine Translation

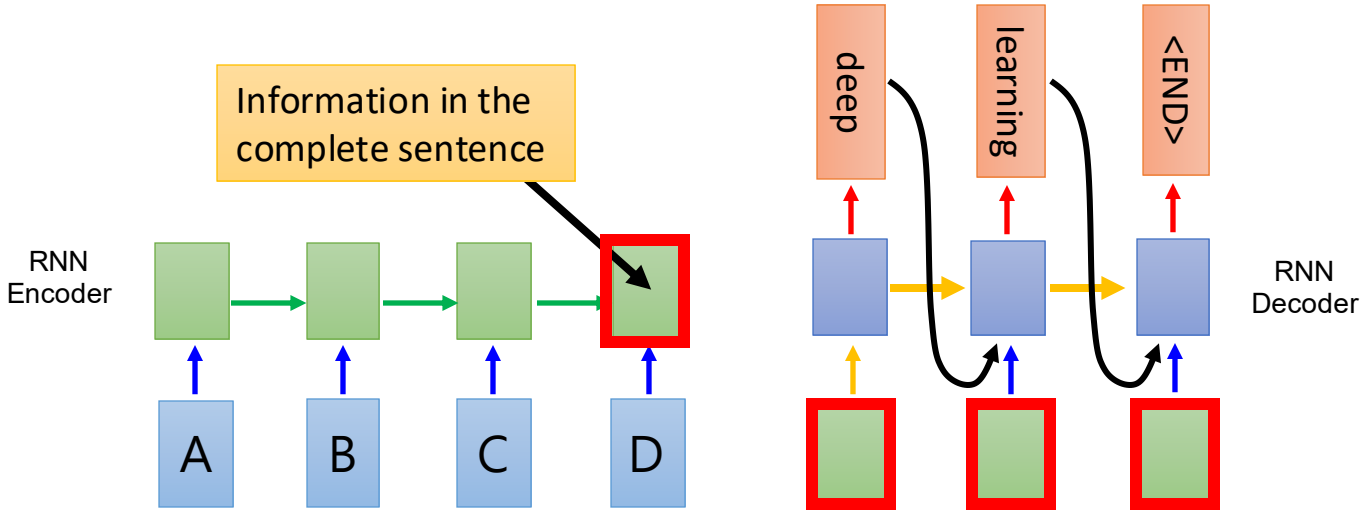
Attention on Sensory Info



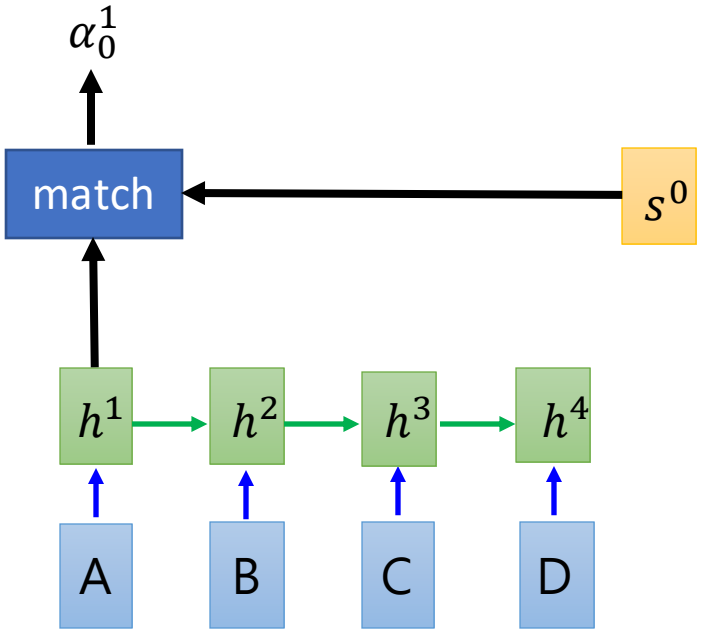
Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.

Machine Translation

- Sequence-to-sequence learning: both input and output are both sequences ***with different lengths***.
- E.g. A B C D → deep learning



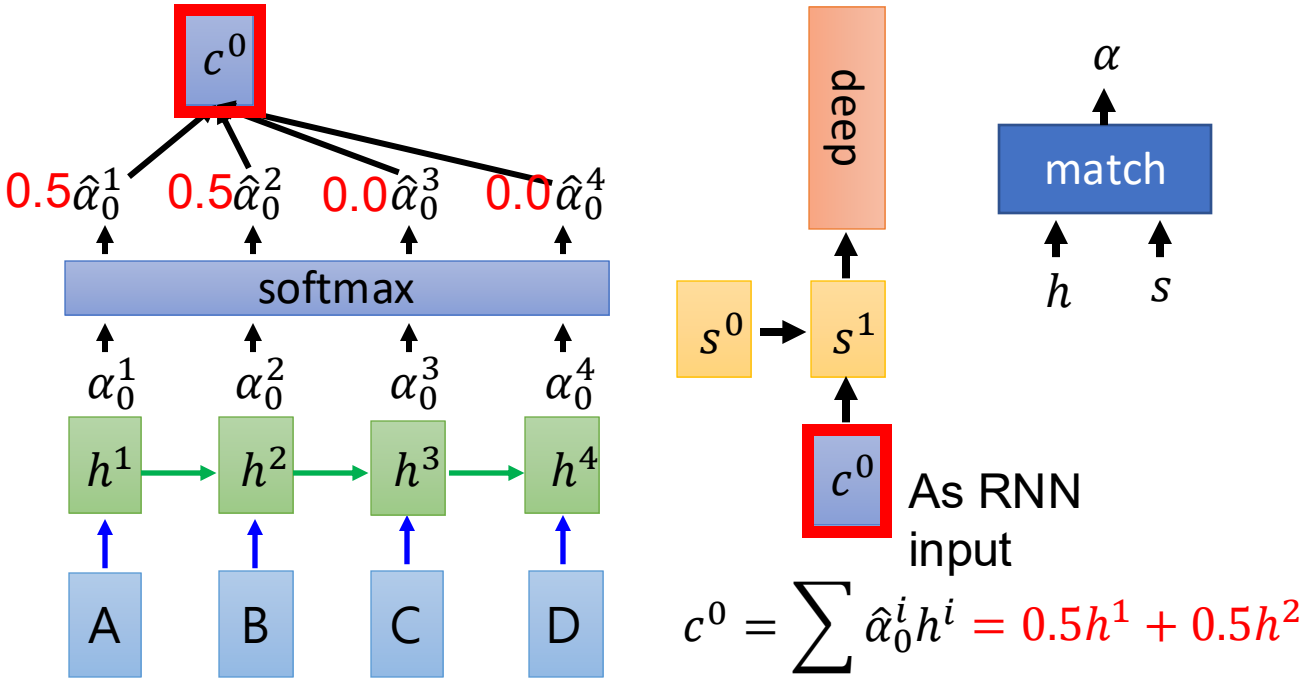
Machine Translation with Attention



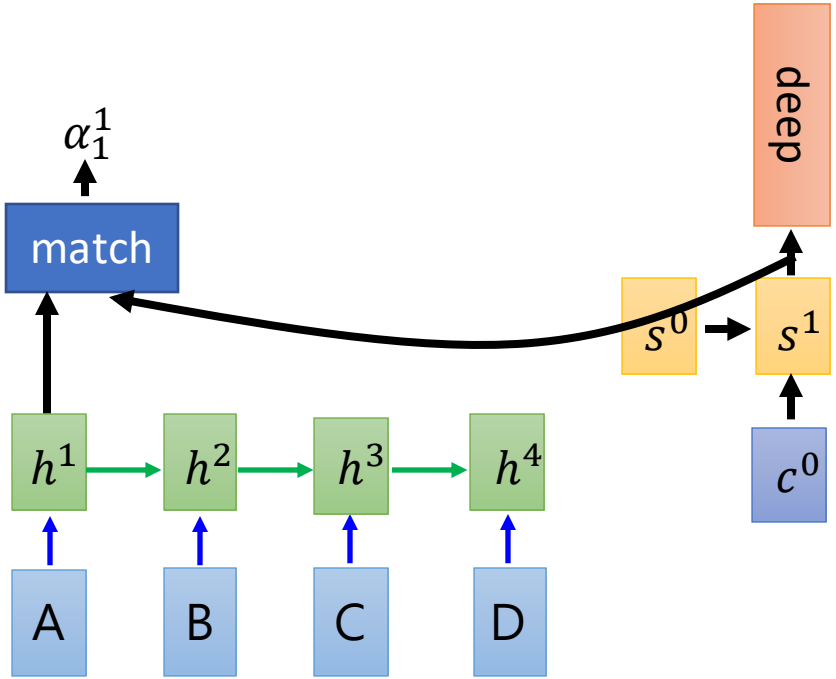
What is **match** ?

- Cosine similarity of s and h
- Small NN whose input is s and h, and output is a scalar
- $\alpha = h^T W s$

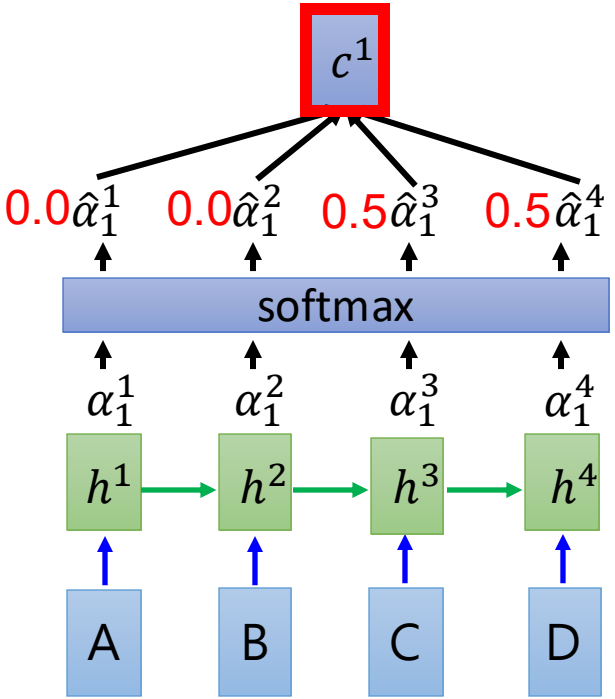
Machine Translation with Attention (cont.)

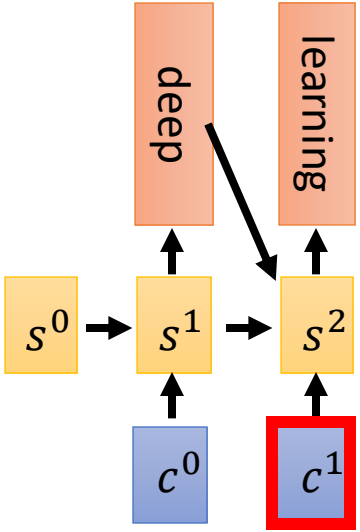


Machine Translation with Attention (cont.)



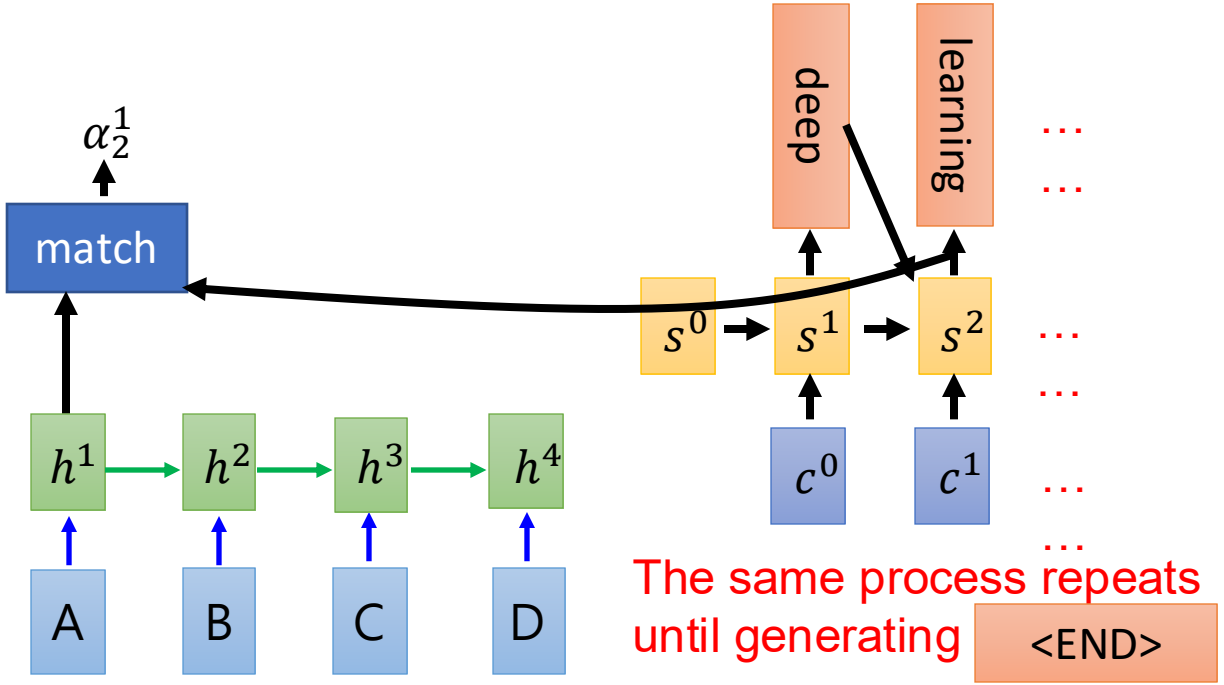
Machine Translation with Attention (cont.)





$$c^1 = \sum \hat{\alpha}_1^i h^i = 0.5h^3 + 0.5h^4$$

Machine Translation with Attention (cont.)



S2S Model for Machine Translation

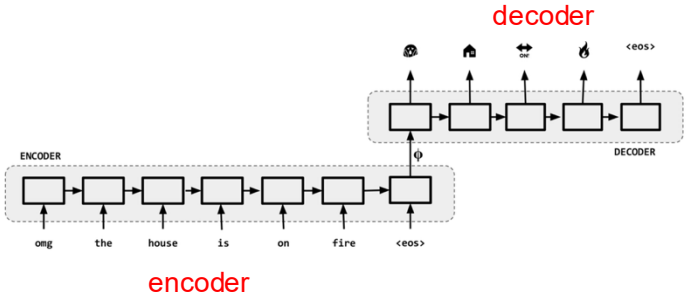
- Input sentence: $\mathbf{x} = (x_1, \cdots, x_{T_x})$
- Encoder RNN:

$$h_t = f(x_t, h_{t-1})$$

$$\phi = q(\{h_1, \cdots, h_{T_x}\})$$

- In S2S models (i.e., Sutskever et al 2014)

$$q(\{h_1, \cdots, h_T\}) = h_T$$

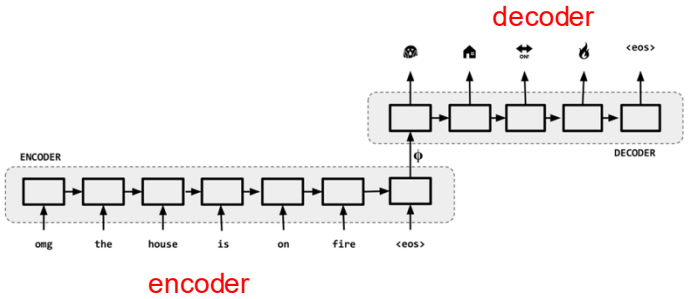


S2S Model for Machine Translation (cont.)

- Target Translation: $\mathbf{y} = (y_1, \cdots, y_{T_y})$
- Decoder RNN:

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t \mid \{y_1, \cdots, y_{t-1}\}, \phi)$$

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, \phi) = g(y_{t-1}, s_t, \phi)$$



Machine Translation with Attention (cont.)

- Encoder is similar.
- Decoder RNN:

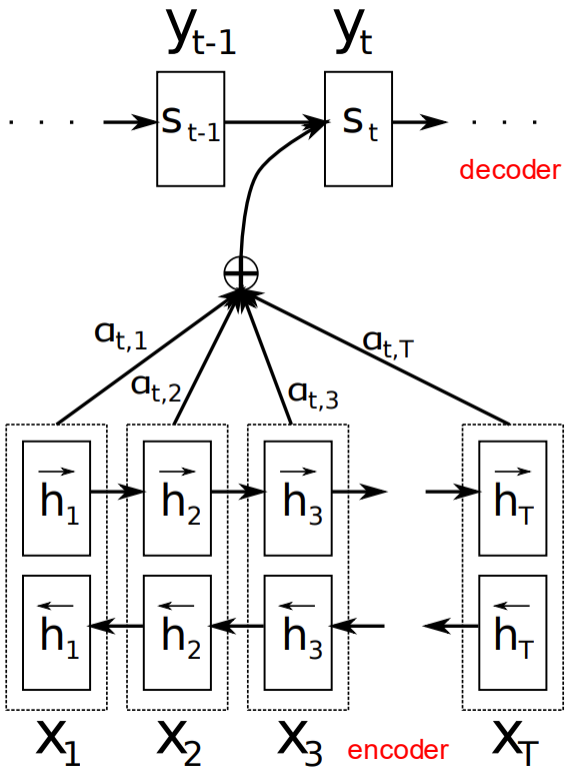
$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1},s_i,c_i)$$

$$s_i = f(s_{i-1},y_{i-1},c_i)$$

- The context vector c_i depends on h_j

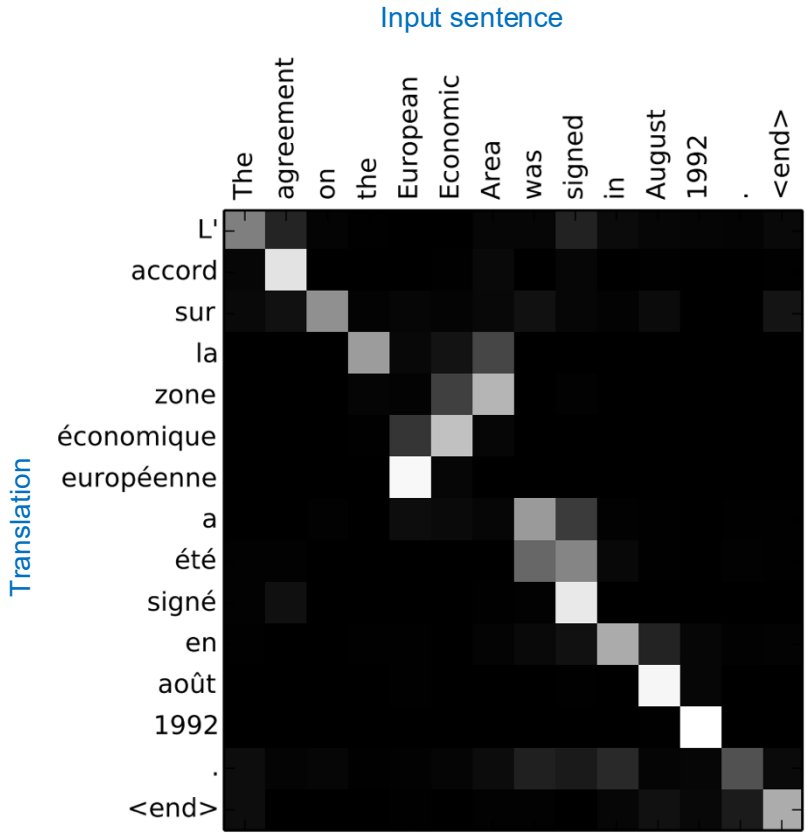
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ik})} \qquad e_{ij} = a(s_{i-1},h_j)$$



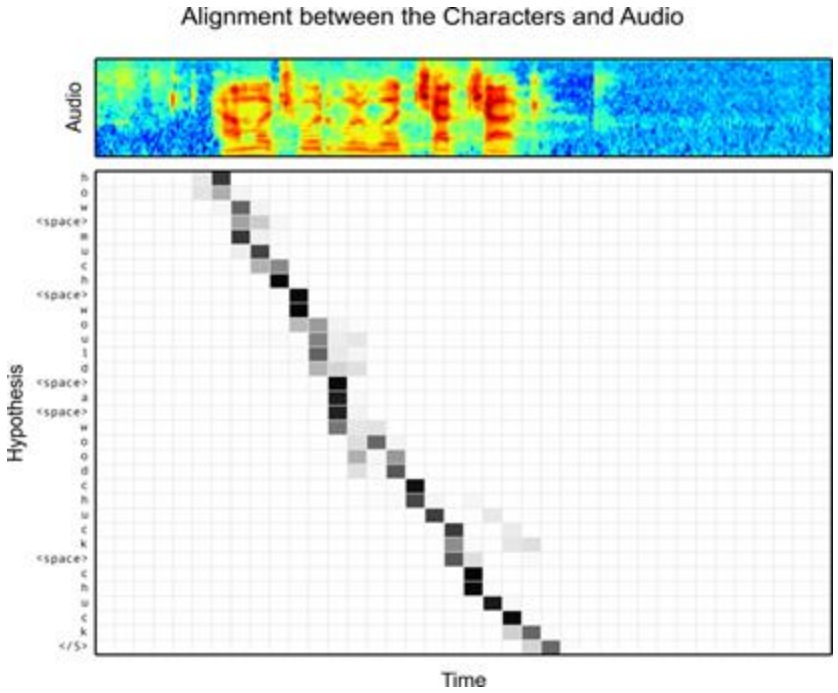
Visualization of Attention for MT

- Attention probability matrix can be useful for visualization.
- Pixels show the weight α_{ij} of the annotation of the j^{th} source word for the i^{th} target word
- Plot is grayscale (0: black, 1: white)



Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.

Speech Recognition with Attention



Chan et al., “Listen, Attend and Spell”, arXiv, 2015 .

Topics for Today

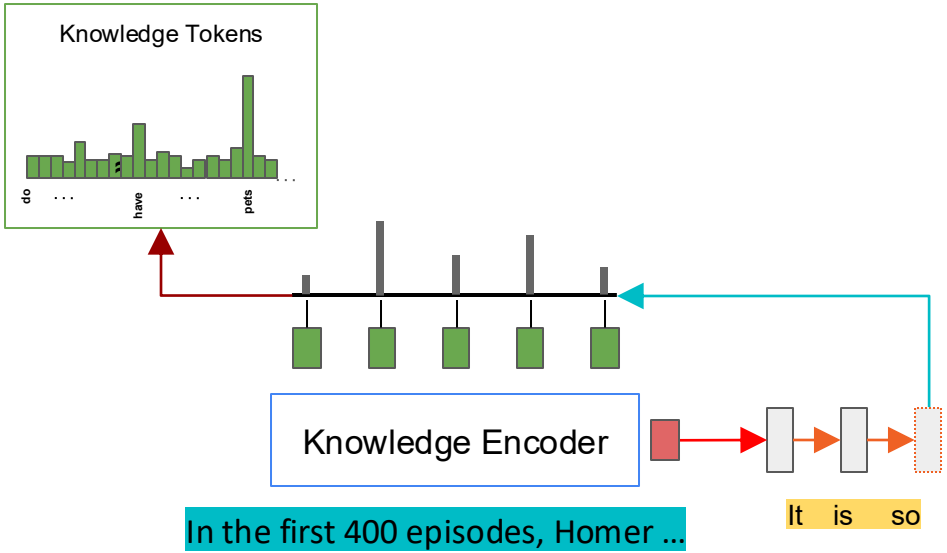
- Encoder-Decoder Architecture
- Beam Search
- Case Study: Response generation in E2E Dialogue Systems
- Limitations of S2S models & Solutions
- Attention
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

Challenge: Previously Unseen Words

- The output will always be limited to the vocabulary of the training data.
- However, knowledge and conversation context can include new words.

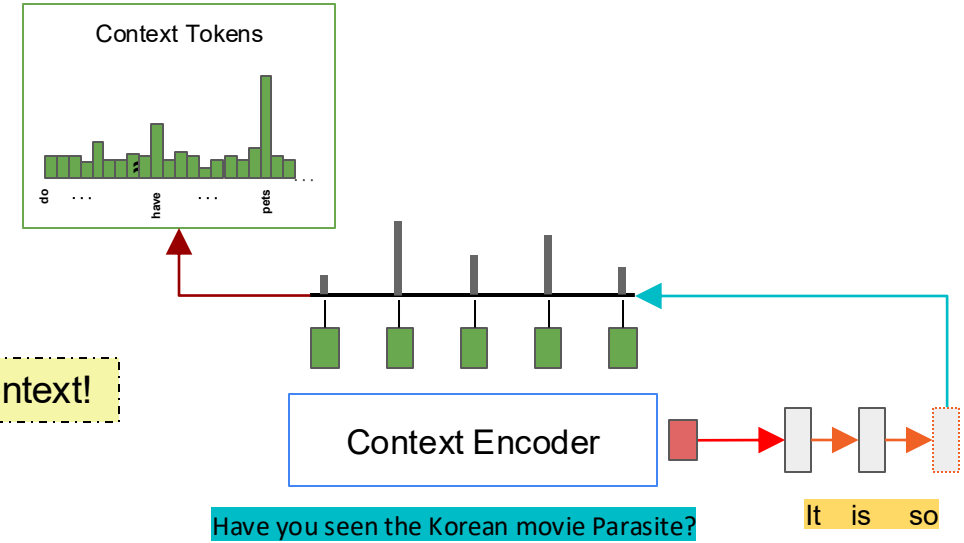
- How to generate new words in conversations?

Copy Mechanism



Get To The Point: Summarization with Pointer-Generator Networks, See et al, ACL 2017

Copy Mechanism

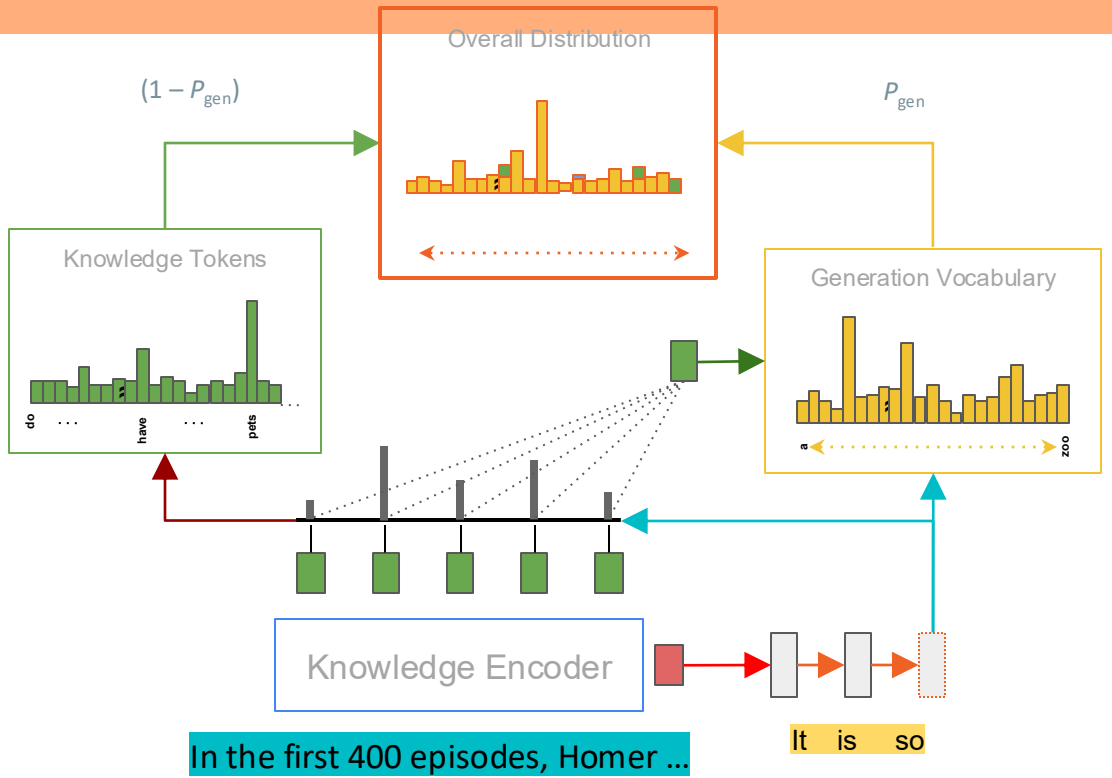


Get To The Point: Summarization with Pointer-Generator Networks, See et al, ACL 2017

Pointer Generator Networks

The overall distribution is obtained by interpolating the LLM output probabilities with input tokens, weighted according to attention.

Out-of-vocabulary (OOV) words can be decoded if they were seen in knowledge.



Get To The Point: Summarization with Pointer-Generator Networks, See et al, ACL 2017

S. Yavuz, et al DeepCopy: Grounded Response Generation with Hierarchical Pointer Networks. SigDial 2019.

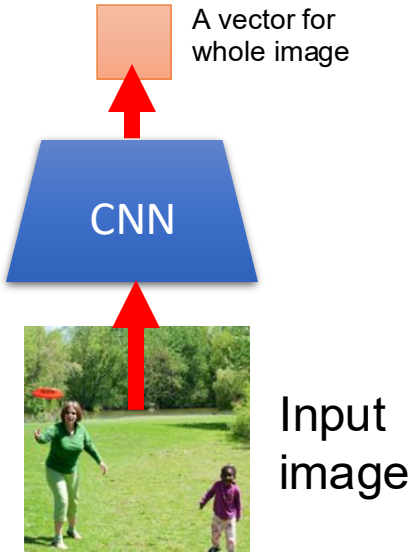
Topics for Thursday

Transformers

- The Transformer Model Architecture
- Self Attention
- Multi-head attention
- The encoder block
- The decoder block

Image Captioning

- Input: image
- Output: word sequence



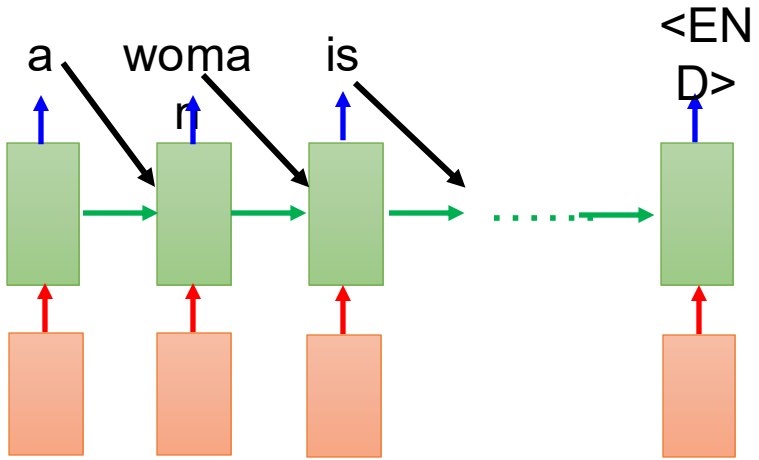
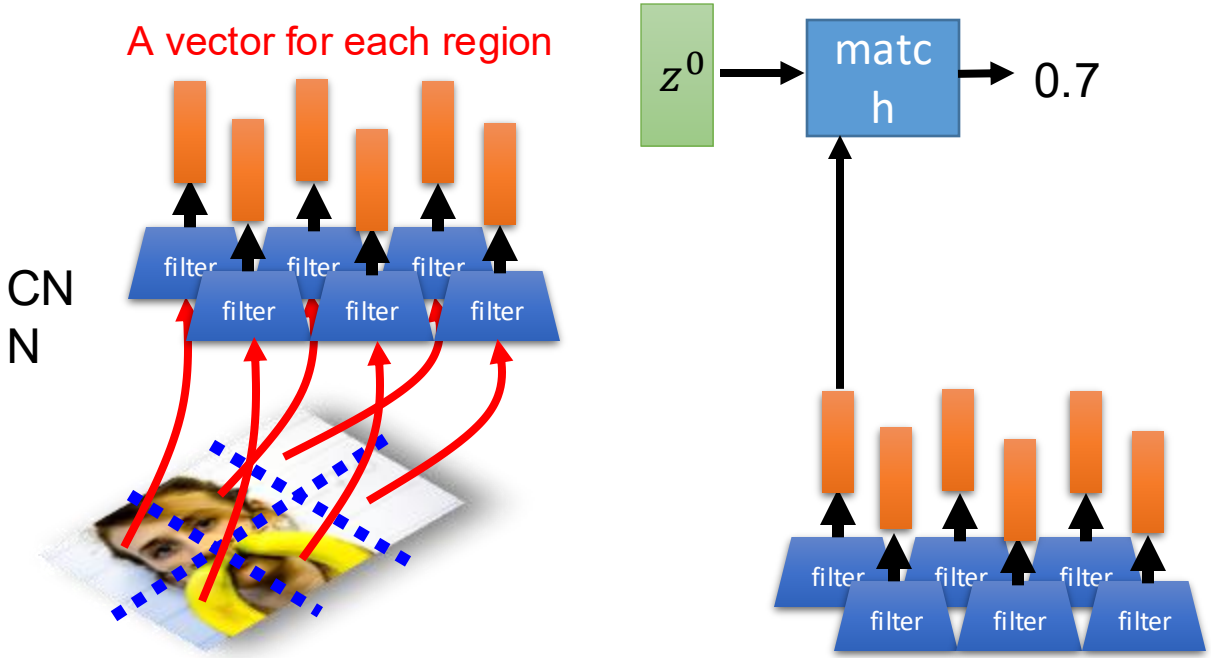


Image Captioning with Attention



Xu et al, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Image Captioning with Attention

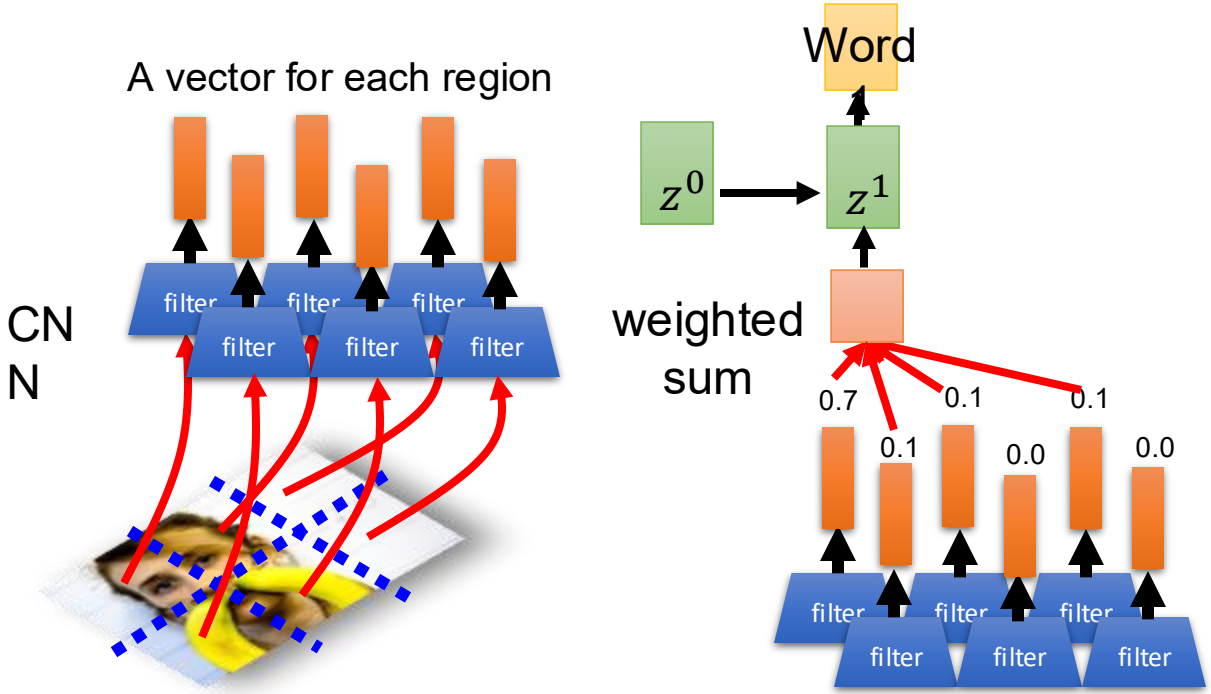


Image Captioning with Attention

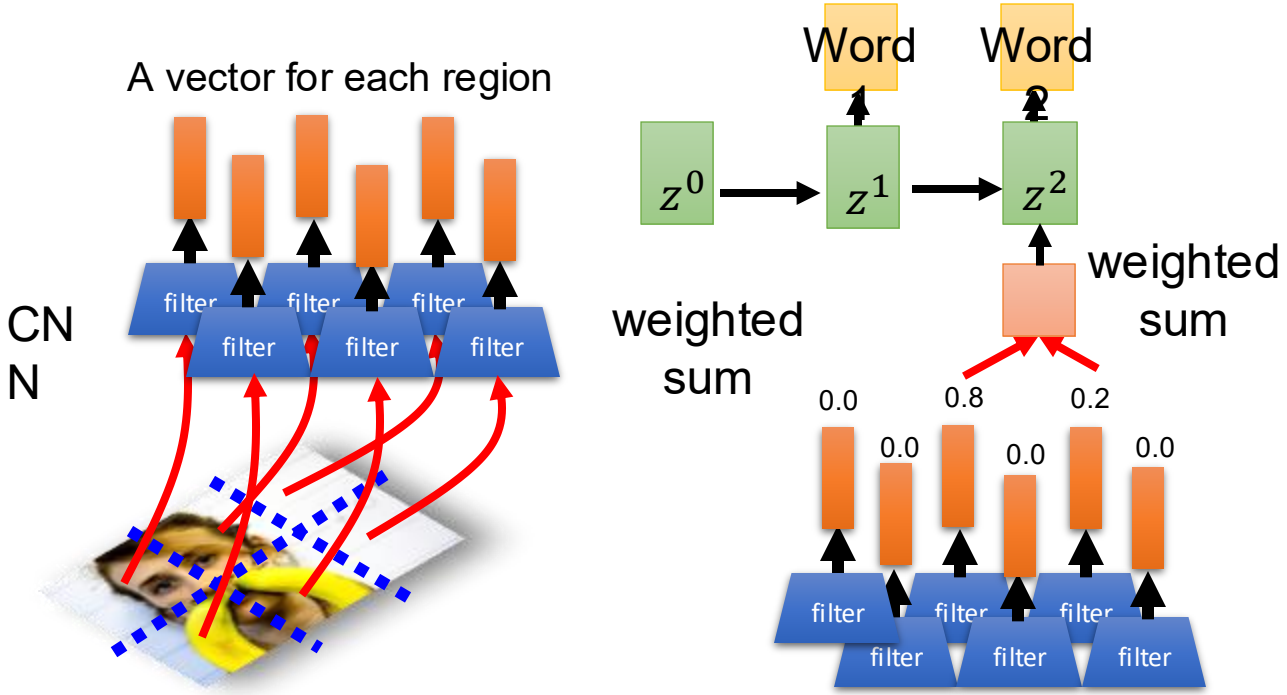


Image Captioning

- Good examples

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Image Captioning

- Bad examples

A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and
a hat on a skateboard.

A person is standing on a beach
with a surfboard.

A woman is sitting at a table
with a large pizza.

A man is talking on his cell phone
while another man watches.

Video Captioning

Ref: A man and a woman ride a motorcycle

A **man** and a **woman** are **talking** on the **road**

Yao et al. 2015, Describing Videos by Exploiting Temporal Structure.

Video Captioning

Ref: A woman is frying food

Someone is **frying** a **fish** in a **pot**

Outline for Today

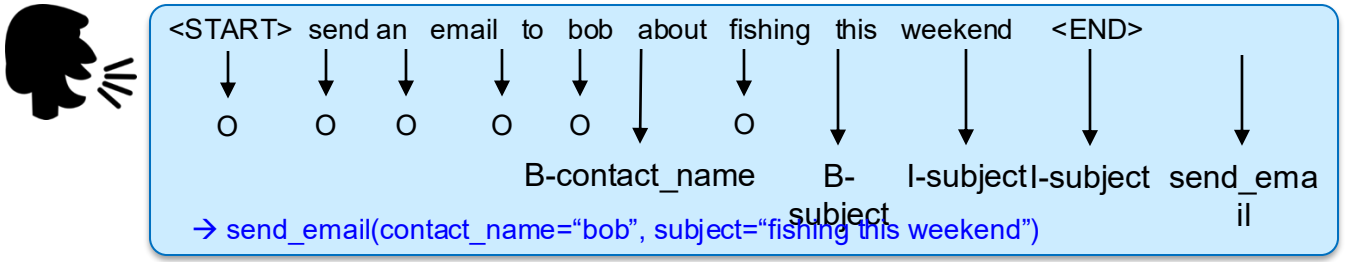
- Attention
- Analysis of attention
- Application of Attention in Tagging
- Attention and Memory Networks
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

- Readings:
  - Chapter 10 of Dive Into Deep Learning (only 10.1 and 10.2)
  - Continuing Chapter 8 of NLP with PyTorch
  - Bahdanu et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.

Slide from earlier lecture

Natural Language Understanding (CLU)

- Tag a word at each timestamp
  - Input: word sequence
  - Output: IOB-format slot tag and intent tag



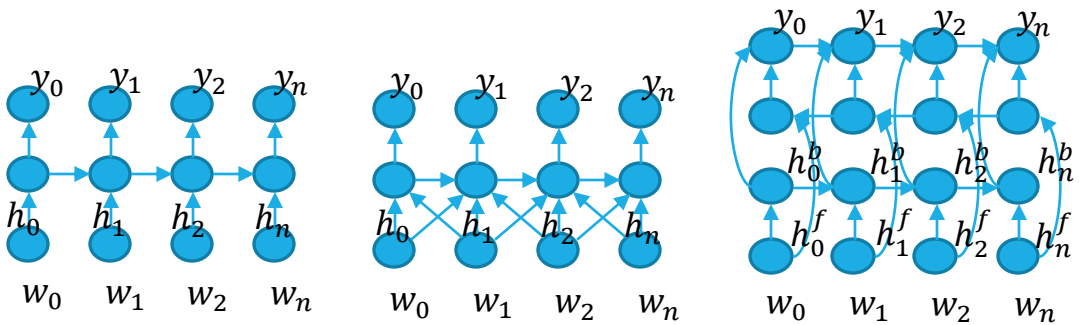
Temporal orders for input and output are the same

Slide from earlier lecture

Recurrent Neural Nets for Slot Tagging – I (Yao et al, 2013; Mesnil et al, 2015)

http://131.107.65.14/en-us/um/people/gzweig/Pubs/Interspeech2013RNNLU.pdf; http://dl.acm.org/citation.cfm?id=2876380

- Variations:
  a. RNNs with LSTM cells
  b. Input, sliding window of n-grams
  c. Bi-directional LSTMs

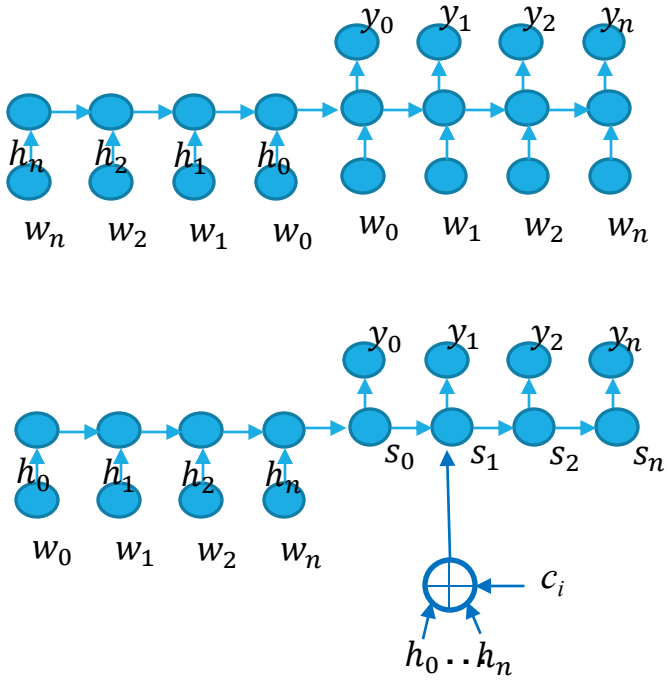


(a) LSTM (b) LSTM-LA (c) bLSTM

Recurrent Neural Nets for Slot Tagging – II (Kurata et al., 2016; Simonnet et al., 2015)

http://www.aclweb.org/anthology/D16-1223; https://hal.archives-ouvertes.fr/hal-01433202

- Encoder-decoder networks
  - Leverages sentence level information
- Attention-based encoder-decoder
  - Use of attention (as in MT) in the encoder-decoder network
  - Attention is estimated using a feed-forward network with input: h_t and s_t at time t

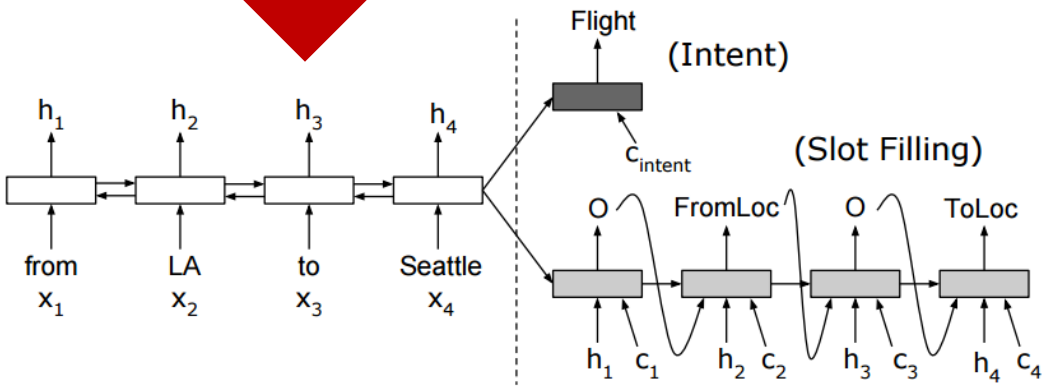


Joint Semantic Frame Parsing

https://arxiv.org/abs/1609.01454

Parallel (Liu and Lane, 2016)

- Intent prediction and slot filling are performed in two branches



- Input sentence is encoded with a biLSTM.
- Decoder estimates tags based on corresponding encoder hidden states as well as attention.
- Comparable performance is achieved due to the attention mechanism.
- The encoder-decoder approach enables multi-task learning with other tasks.

Outline for Today

- Attention
- Analysis of attention
- Application of Attention in Tagging
- Attention and Memory Networks
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

- Readings:
  - Chapter 10 of Dive Into Deep Learning (only 10.1 and 10.2)
  - Continuing Chapter 8 of NLP with PyTorch
  - Bahdanu et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.

Machine Reading Comprehension

- **S**tanford **Qu**estion **A**nswering **D**ataset (SQuAD)

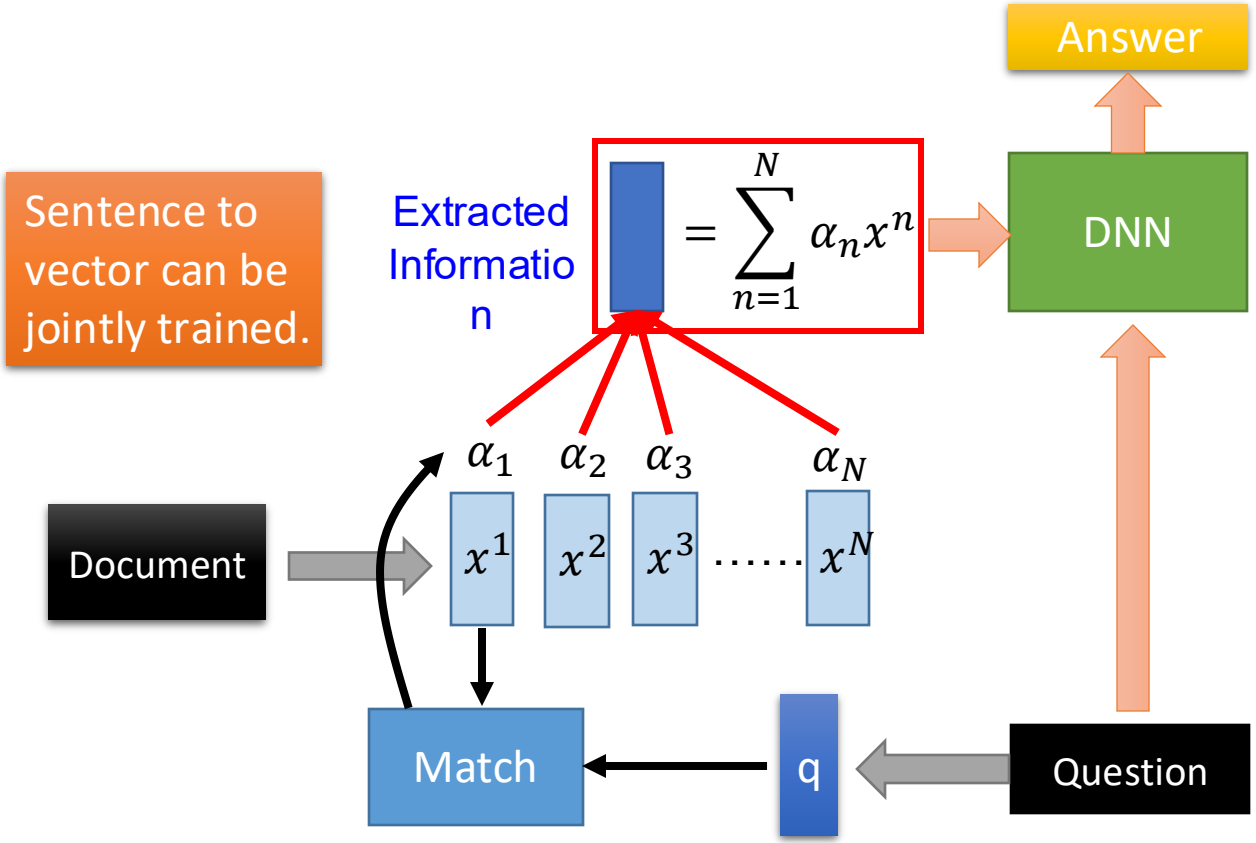
The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

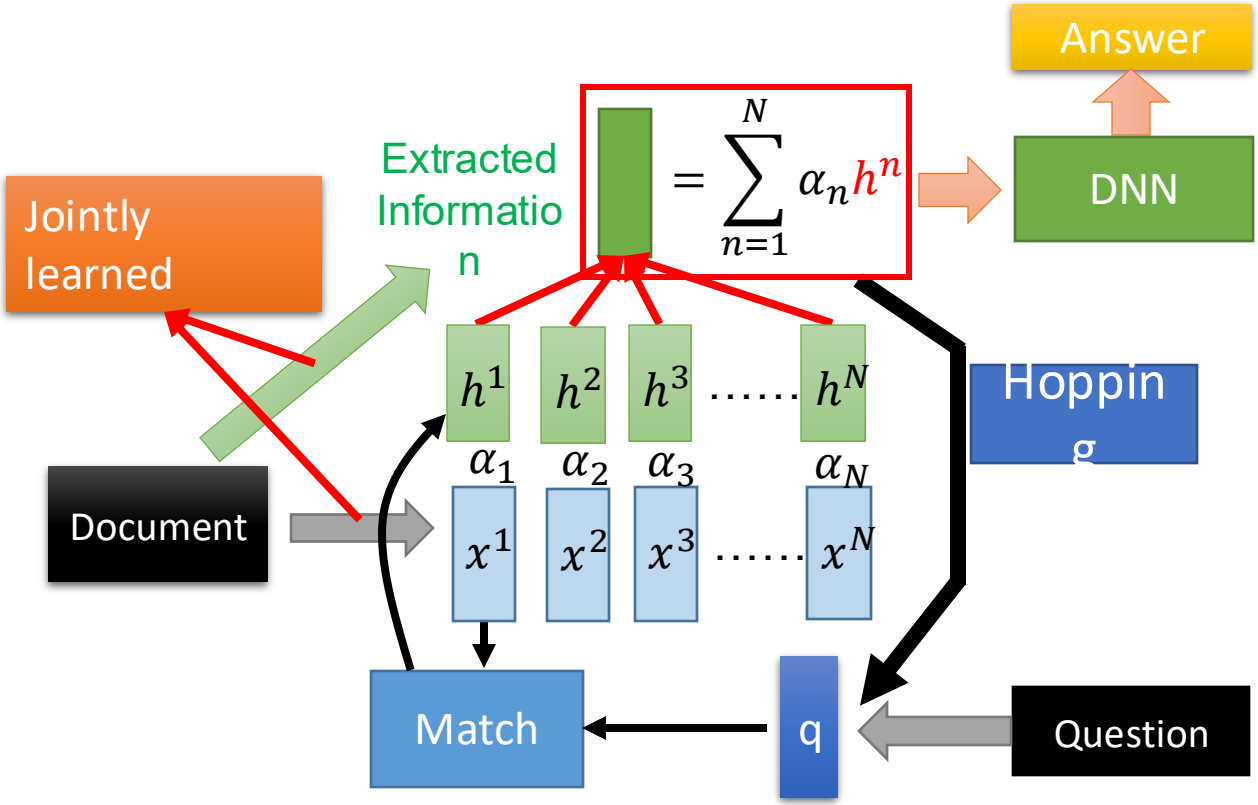
We can use attention to focus on parts of long documents!

Machine Reading Comprehension



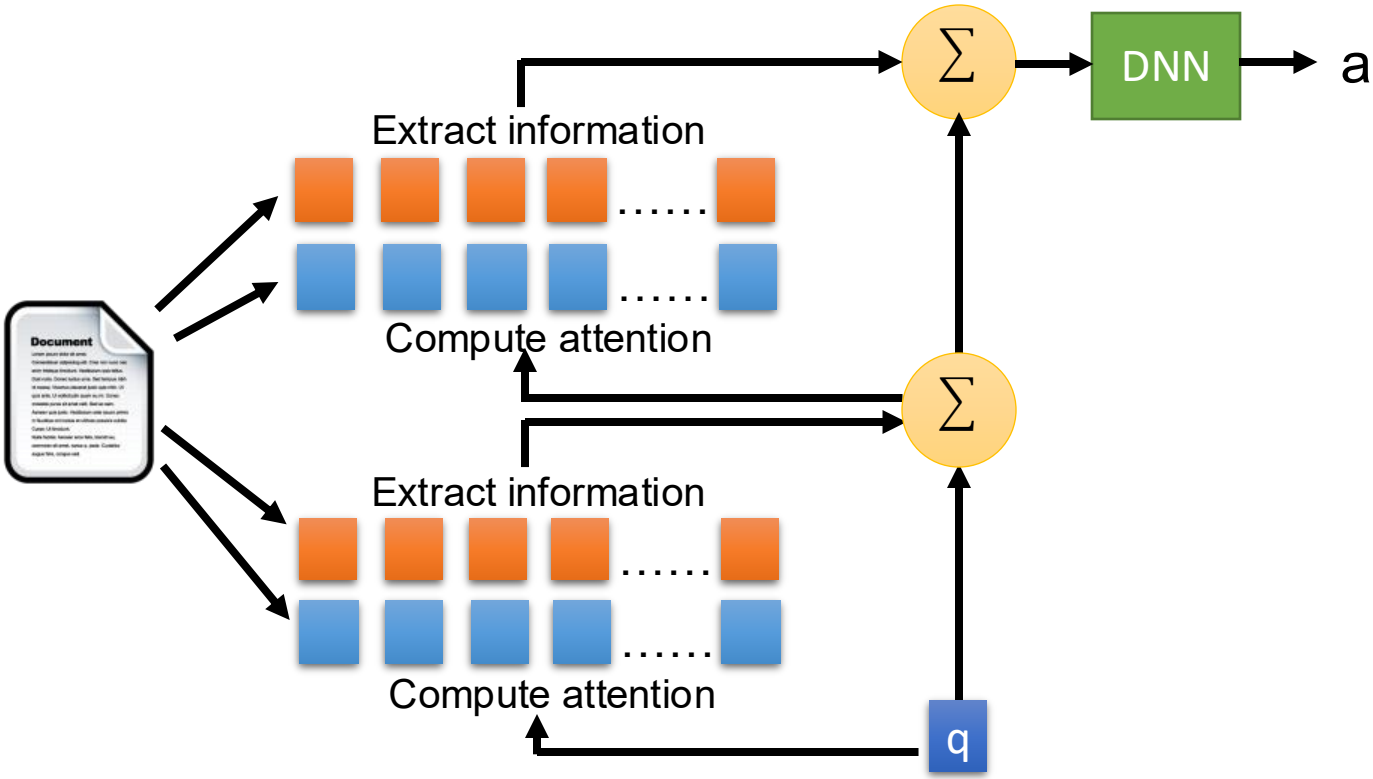
https://arxiv.org/pdf/1704.00051.pdf

Machine Reading Comprehension (cont.)



https://arxiv.org/pdf/1704.00051.pdf

Memory Networks



Memory Networks

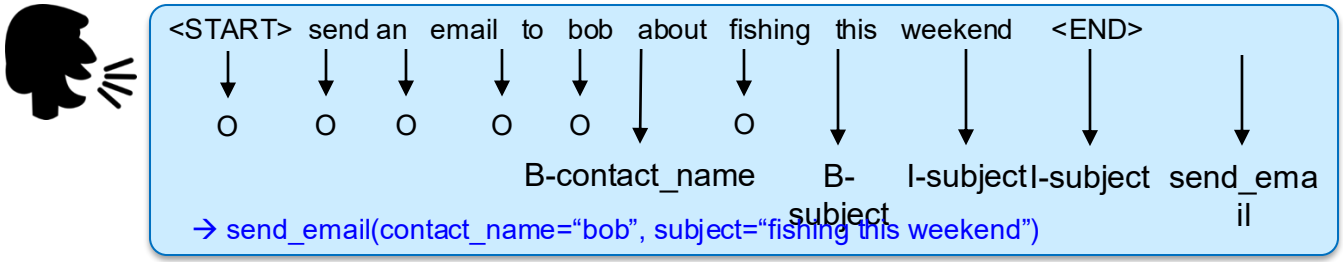
- Multi-hop performance analysis on babI task:

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom		**Prediction: bathroom**		

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow		**Prediction: yellow**		

Memory networks: other applications

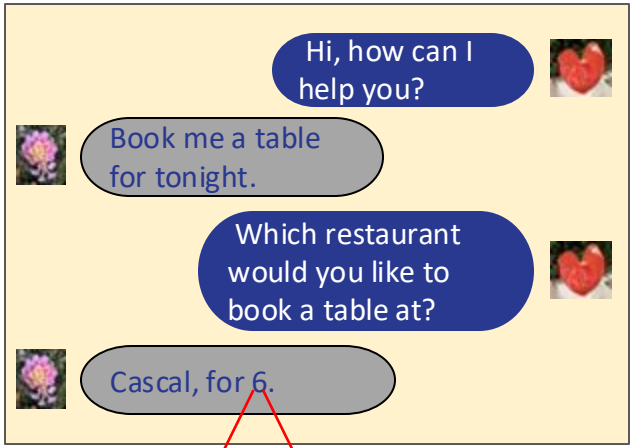
- Natural language understanding in conversational systems mainly consider the last user utterance.

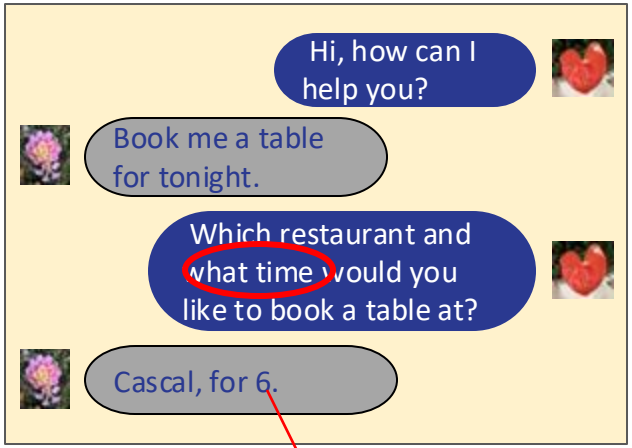


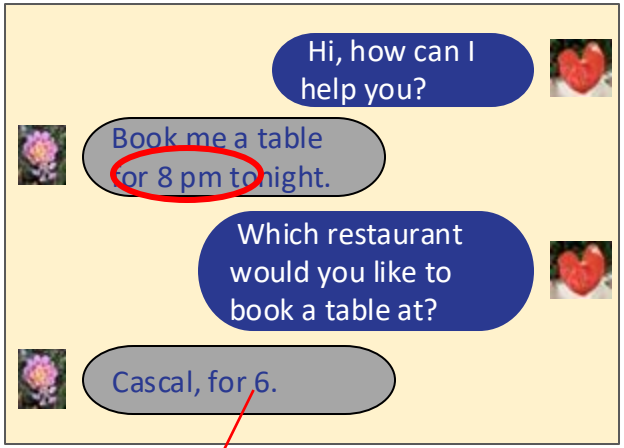
Challenge: Integrating Conversation Context

- User utterances are highly ambiguous in isolation

Restaurant booking

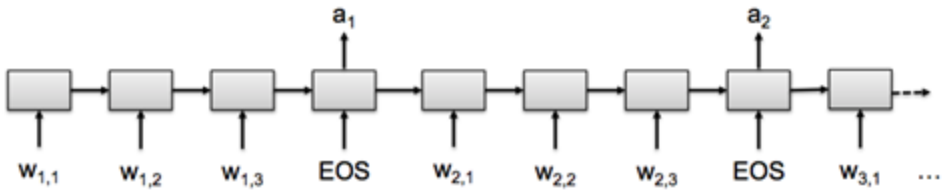






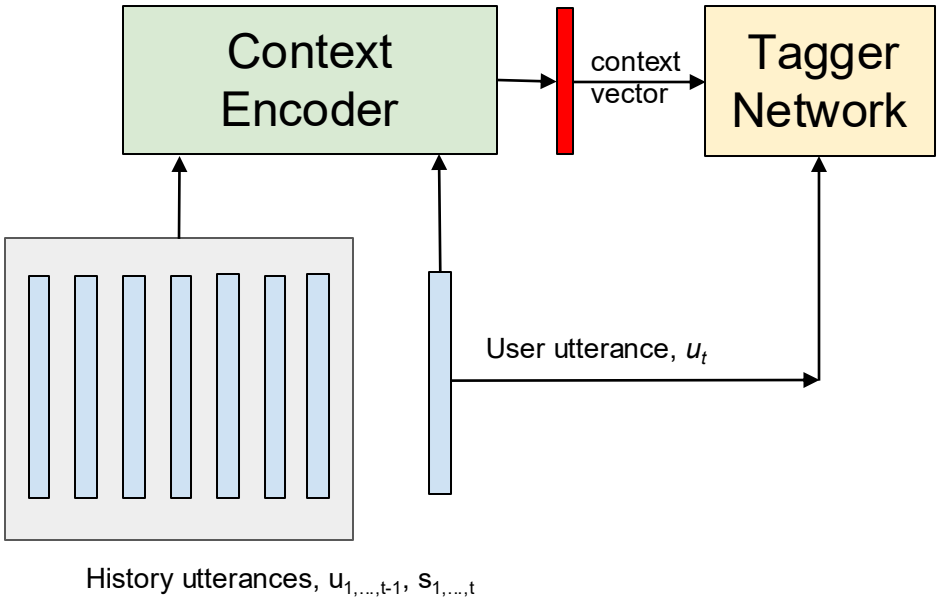
Previous Work on Contextual CLU

- Seq2Seq model (Hori et al, 2015)
  - Words are input one at a time, tags are output at the end of each uttera



- Extension: LSTM with speaker role dependent layers

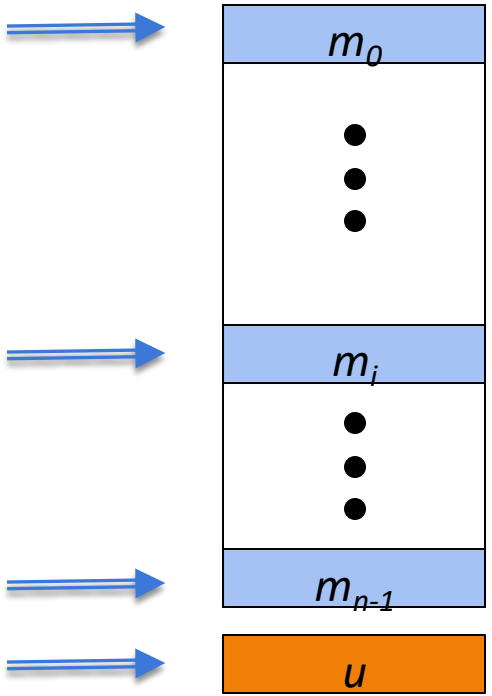
Use of Conversation Context for CLU



- Encoding current and previous conversation turns
- Combining history embeddings
- Using them during CLU for the current conversation turn

Use of Conversation Context for CLU (cont.)

U: "i d like to purchase tickets to see deepwater horizon"
S: "for which theatre"
U: "angelika"
S: "you want them for angelika theatre?"
U: "yes angelika"
S: "how many tickets would you like ?"
U: "3 tickets for saturday"
S: "What time would you like ?"
U: "Any time on saturday is fine"
S: "okay , there is 4:10 pm , 5:40 pm and 9:20 pm"
U: "Let's do 5:40"



E2E MemNN for Contextual CLU

Chen et al., End-to-End Memory Networks with Knowledge Carryover for Multi-Turn Spoken Language Understanding, Interspeech 2016.

1. Sentence Encoding

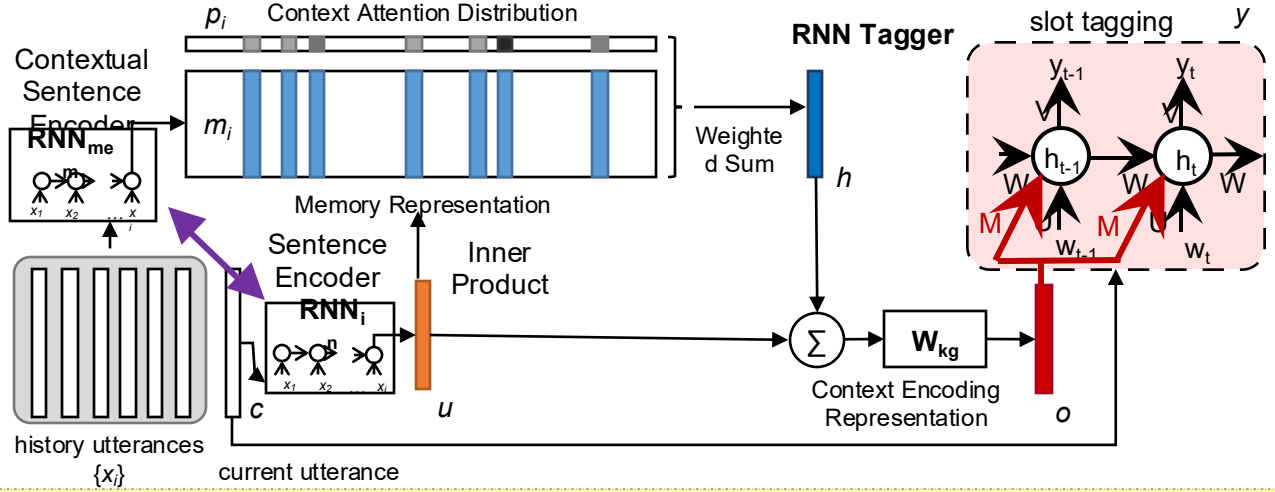
$$m_i = \text{RNN}_{\text{mem}}(x_i)$$
$$u = \text{RNN}_{\text{in}}(c)$$

2. Context Attention

$$p_i = \text{softmax}(u^T m_i)$$

3. Context Encoding

$$h = \sum_i p_i m_i \quad o = W_{\text{kg}}(h + u)$$



Idea: additionally incorporating contextual knowledge during slot tagging

→ track dialogue states in a latent way

Analysis of Attention

U: "i d like to purchase tickets to see deepwater horizon" → 0.69

S: "for which theatre"

U: "angelika"

S: "you want them for angelika theatre?"

U: "yes angelika"

S: "how many tickets would you like ?"

U: "3 tickets for saturday"

S: "What time would you like ?" → 0.13

U: "Any time on saturday is fine"

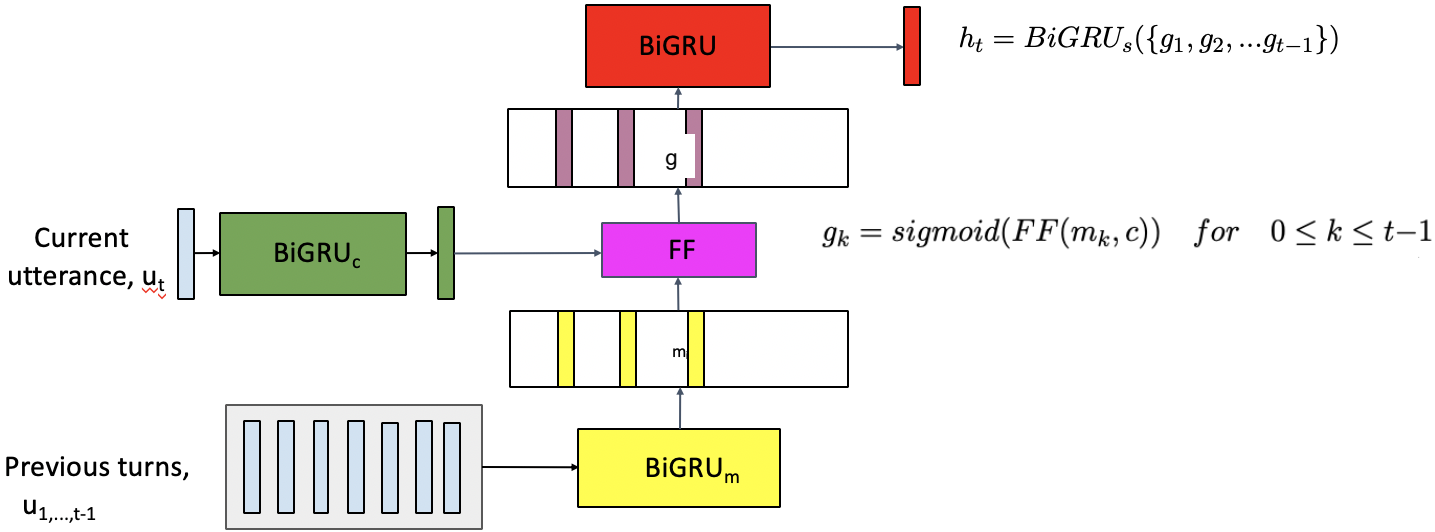
S: "okay , there is 4:10 pm , 5:40 pm and 9:20 pm" → 0.16

U: "Let's do 5:40"

Sequential Dialogue Encoder (SDEN)

Bapna et al, Sequential Dialogue Context Modeling for Spoken Language Understanding, SigDial, 2017.

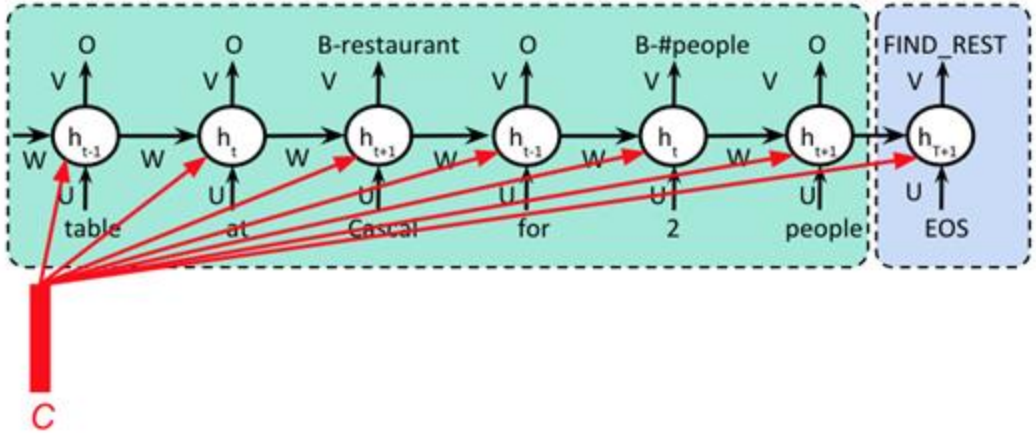
- E2E MemNN for Contextual CLU relies on cosine similarity for attention.
- Can we use a machine learning model instead?



Sequential Dialogue Encoder (SDEN)

Bapna et al, Sequential Dialogue Context Modeling for Spoken Language Understanding, SigDial, 2017.

- E2E MemNN for Contextual CLU relies on cosine similarity for attention.
- Can we use a machine learning model instead?



Outline for Today

- Attention
- Analysis of attention
- Application of Attention in Tagging
- Attention and Memory Networks
- Attention to Implement Copy Mechanisms (and Pointer Generator Networks)

- Readings:
  - Chapter 10 of Dive Into Deep Learning (only 10.1 and 10.2)
  - Continuing Chapter 8 of NLP with PyTorch
  - Bahdanu et al., Neural Machine Translation by Jointly Learning to Align and Translate. ICLR, 2015.

Grounding Conversations in Knowledge

- Sequence-to-sequence models typically generate generic responses.
- Information previously unseen in conversations may be available elsewhere
  - Unstructured knowledge: news articles, Wikipedia, reviews, etc.
  - Structured knowledge: knowledge graphs

User: Can you find me an Italian restaurant in Santa Clara?

System: Il Fornaio is an Italian restaurant in Santa Clara. Their **gnocchi gorgonzola** and **spaghetti carbonara** are quite popular according to reviews.

"Our favorites: caesar salad, gamberini in padella, risotto, **spaghetti carbonara,** and the arrabiata pasta." in 11 reviews

"That all being said, the **Gnocchi Gorgonzola** was divine, and much lighter than I expected." in 6 reviews

Grounding Conversations in Knowledge (cont.)

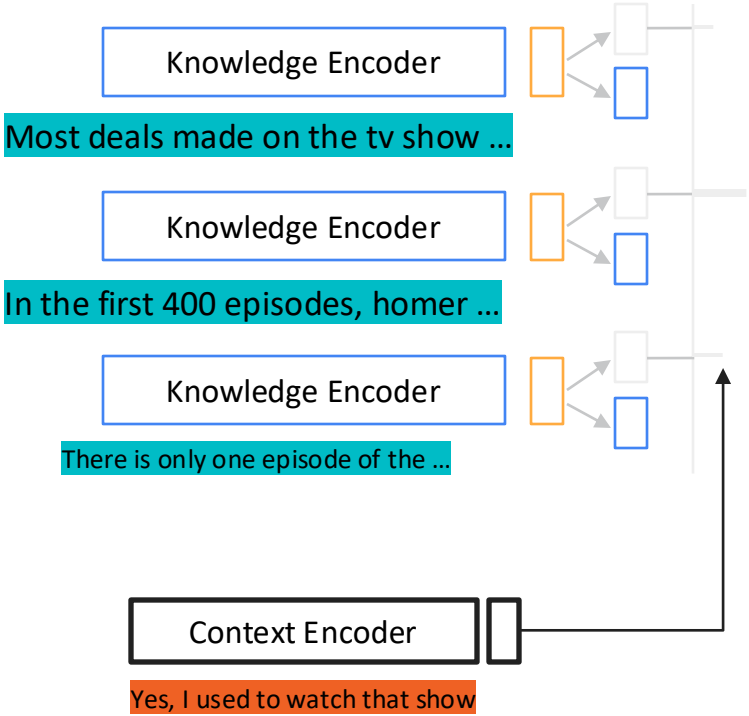
- A simple approach is to have a network for **knowledge selection**.
- Given a conversation context, c_t, and a set of knowledge sentences $k_1,\ldots,k_{|K|}$
  - $\underline{k}_t = \text{Argmax}_{j=1,\ldots,|K|} f(c_t, k_j)$
  - f() could be similarity measure or a network that could be trained.
- Then, feed encodings of both c_t and $\underline{k}_t$ as input to the decoder.

Gopalakrishnan et al., Interspeech 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversa

- Knowledge selection could be trained jointly with response generation.

Dinan et al, ICLR 2019. Wizard of Wikipedia: Knowledge-Powered Conversational agents

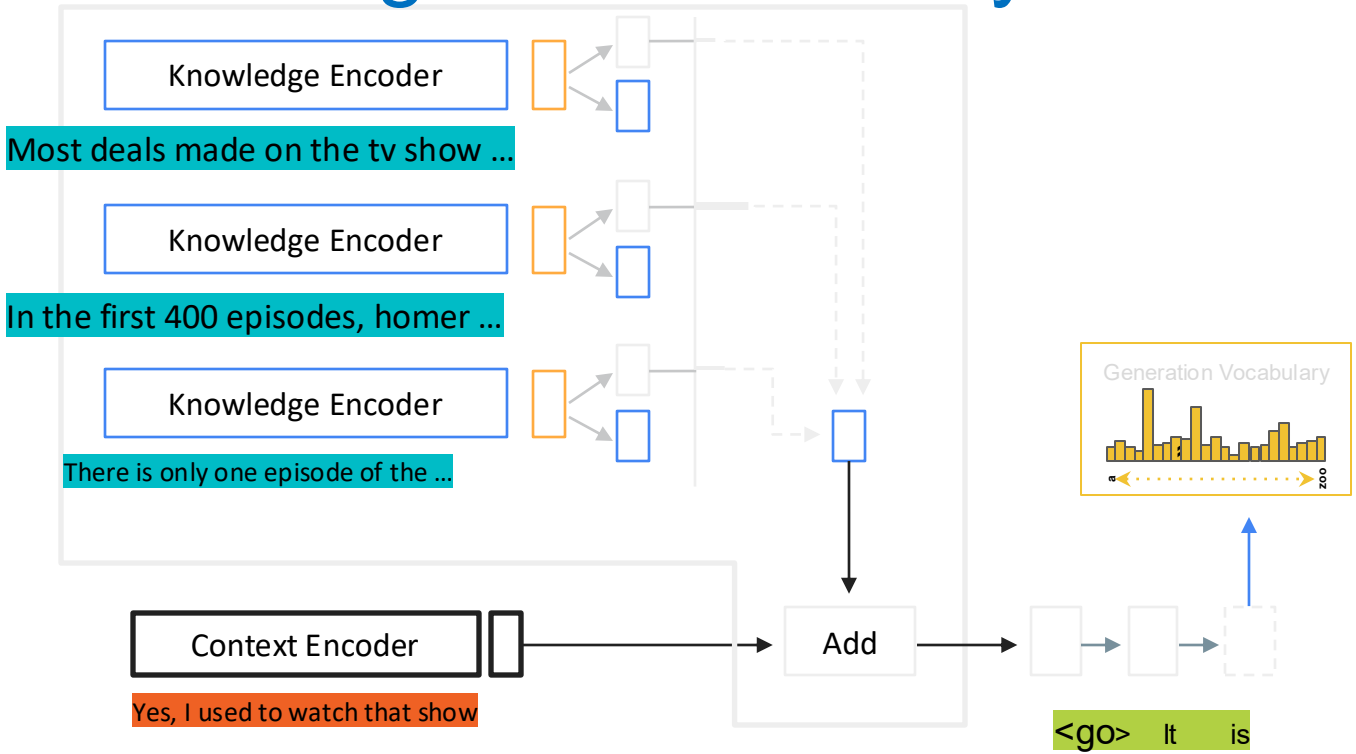
Grounding Conversations in Knowledge: E2E Memory Networks



- All knowledge snippets are encoded in a representation using conversation context.

Ghazvininejad, AAAI 2018. A Knowledge-Grounded Neural Conversation Model.

Grounding Conversations in Knowledge: E2E Memory Networks



- All knowledge snippets are encoded in a representation using conversation context.
- The knowledge and context representation are then used to initialize the decoder.