CS388: Natural Language Processing
Lecture 19: Reading Comprehension

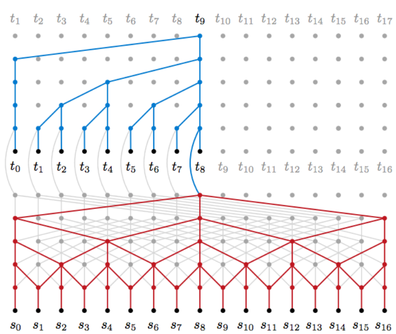
Greg Durrett

Administrivia

- Project 2 due Friday at 5pm
- Project proposals due next Thursday
  - Spec posted on course website — I'll pitch some ideas/interesting papers from EMNLP on Tuesday

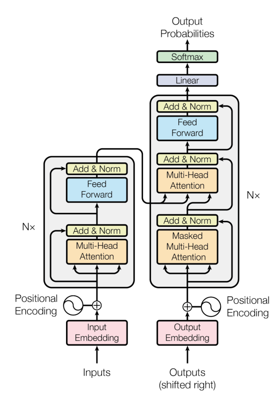
Recall: CNNs for Machine Translation

- "ByteNet": operates over characters (bytes)
- Encode source sequence w/dilated convolutions
- Predict *n*th target character by looking at the *n*th position in the source and a dilated convolution over the *n*-1 target tokens so far
- To deal with divergent lengths, t_n actually looks at $s_{n\alpha}$ where α is a heuristically-chosen parameter
- Assumes mostly monotonic translation



Kalchbrenner et al. (2016)

Recall: Transformers



- Encoder and decoder are both transformers
- Decoder consumes the previous generated token (and attends to input), but has *no recurrent state*

Vaswani et al. (2017)

This Lecture

- Types of question answering/reading comprehension
- Memory networks
- CNN/Daily Mail task: Attentive Reader
- SQuAD task: Bidirectional Attention Flow

Reading Comprehension

Classical Question Answering

- Form semantic representation from semantic parsing, execute against structured knowledge base

Q: "where was Barack Obama born"

```
λx. type(x, Location) ∧ born_in(Barack_Obama, x)
```

(other representations like SQL possible too…)

- How to deal with open-domain data/relations? Need data to learn how to ground every predicate or need to be able to produce predicates in a zero-shot way

QA from Open IE

(a) **CCG parse** builds an underspecified semantic representation of the sentence.

Former	municipalities	in	Brandenburgh
N/N	N	$N\backslash N/NP$	NP
$\lambda f \lambda x. f(x) \land former(x)$	$\lambda x. municipalities(x)$	$\lambda f \lambda x \lambda y. f(y) \land in(y, x)$	$Brandenburg$

N: $\lambda x. former(x) \land municipalities(x)$ $>$

$N\backslash N$: $\lambda f \lambda y. f(y) \land in(y, Brandenburg)$ $>$

N: $l_0 = \lambda x. former(x) \land municipalities(x) \land in(x, Brandenburg)$ $<$

(b) **Constant matches** replace underspecified constants with Freebase concepts

$l_0 = \lambda x. former(x) \land municipalities(x) \land in(x, Brandenburg)$

$l_1 = \lambda x. former(x) \land municipalities(x) \land in(x, \mathtt{Brandenburg})$

$l_2 = \lambda x. former(x) \land municipalities(x) \land \mathtt{location.containedby}(x, \mathtt{Brandenburg})$

$l_3 = \lambda x. former(x) \land \mathtt{OpenRel}(x, \mathtt{Municipality}) \land \mathtt{location.containedby}(x, \mathtt{Brandenburg})$

$l_4 = \lambda x. \mathtt{OpenType}(x) \land \boxed{\mathtt{OpenRel}(x, \mathtt{Municipality})} \land \mathtt{location.containedby}(x, \mathtt{Brandenburg})$

- Why use the KB at all? Why not answer questions directly from text? Like information retrieval!

Choi et al. (2015)

What can't KB QA systems do?

- What were the main causes of World War II? — requires summarization
- Can you get the flu from a flu shot? — want IR to provide an explanation of the answer
- What temperature should I cook chicken to? — could be written down in a KB but probably isn't
- Today: can we do QA when it requires retrieving the answer from a passage?

Reading Comprehension

- "AI challenge problem": answer question given context
- Recognizing Textual Entailment (2006)
- MCTest (2013): 500 passages, 4 questions per passage
- Two questions per passage explicitly require cross-sentence reasoning

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

3) Where did James go after he went to the grocery store?
A) his deck
B) his freezer
C) a fast food restaurant
D) his room

Richardson (2013)

Baselines

- N-gram matching: append question + each answer, return answer which gives highest n-gram overlap with a sentence
- Parsing: find direct object of "pulled" in the document where the subject is James

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

2) What did James pull off of the shelves in the grocery store?
A) pudding
B) fries
C) food
D) splinters

- Don't need any complex semantic representations

Richardson (2013)

Reading Comprehension

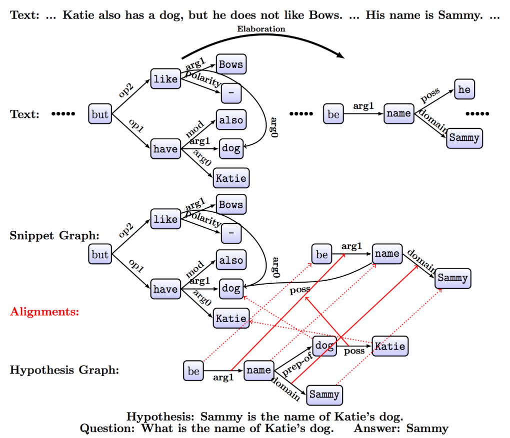
ngram sliding window

	MC160 Test	MC500 Test
Baseline (SW+D)	66.25	56.67
RTE	59.79‡	53.52
Combined	67.60	60.83‡

- Classic textual entailment systems don't work as well as n-grams
- Scores are low partially due to questions spanning multiple sentences
- Unfortunately not much data to train better methods on (2000 questions)

Richardson (2013)

MCTest State of the Art



- Match an AMR (abstract meaning representation) of the question against the original text
- 70% accuracy (roughly 10% better than baseline)

Sachan and Xing (2016)

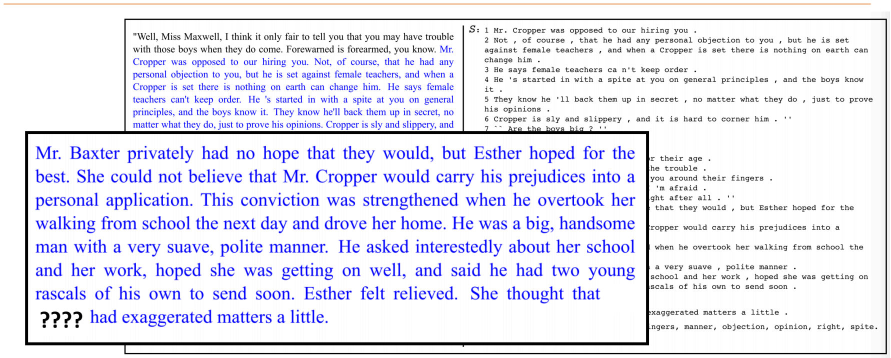
Dataset Explosion

- 10+ QA datasets released since 2015
  - Children's Book Test, CNN/Daily Mail, SQuAD, TriviaQA are most well-known (others: SearchQA, MS Marco, RACE, WikiHop, …)
- Question answering: questions are in natural language
  - Answers: multiple choice or require picking from the passage
  - Require human annotation
- "Cloze" task: word (often an entity) is removed from a sentence
  - Answers: multiple choice, pick from passage, or pick from vocabulary
  - Can be created automatically from things that aren't questions

Dataset Properties

- Axis 1: QA vs. cloze
- Axis 2: single-sentence vs. passage
  - Often shallow methods work well because most answers are in a single sentence (SQuAD, MCTest)
  - Some explicitly require linking between multiple sentences (MCTest)
- Axis 3: single-document (datasets in this lecture) vs. multi-document (TriviaQA, WikiHop, HotPotQA, …)

Children's Book Test

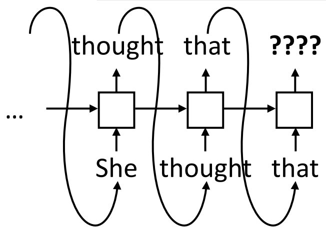


- Children's Book Test: take a section of a children's story, block out an entity and predict it (one-doc multi-sentence cloze task)

Hill et al. (2015)

LSTM Language Models

Mr. Baxter privately had no hope that they would, but Esther hoped for the best. She could not believe that Mr. Cropper would carry his prejudices into a personal application. This conviction was strengthened when he overtook her walking from school the next day and drove her home. He was a big, handsome man with a very suave, polite manner. He asked interestedly about her school and her work, hoped she was getting on well, and said he had two young rascals of his own to send soon. Esther felt relieved. She thought that **????** had exaggerated matters a little.



- Predict next word with LSTM LM
- Context: either just the current sentence (query) or the whole document up to this point (query+context)

Hill et al. (2015)

Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

Methods	Named Entities
Humans (query)(*)	0.520
Humans (context+query)(*)	***0.816***
Maximum frequency (corpus)	0.120
Maximum frequency (context)	0.335
Sliding window	0.168
Word distance model	0.398
Kneser-Ney language model	0.390
Kneser-Ney language model + cache	0.439
LSTMs (query)	0.408
LSTMs (context+query)	0.418

- Neural LMs aren't better than n-gram LMs

Hill et al. (2015)

Children's Book Test: Results

- Present 10 options drawn from the text (correct + 9 distractors), ask the model to pick among them

Methods	Named Entities	Common Nouns	Verbs	Prepositions
Humans (query)(*)	0.520	0.644	0.716	0.676
Humans (context+query)(*)	***0.816***	***0.816***	***0.828***	0.708
Maximum frequency (corpus)	0.120	0.158	0.373	0.315
Maximum frequency (context)	0.335	0.281	0.285	0.275
Sliding window	0.168	0.196	0.182	0.101
Word distance model	0.398	0.364	0.380	0.237
Kneser-Ney language model	0.390	0.544	0.778	0.768
Kneser-Ney language model + cache	0.439	0.577	0.772	0.679
LSTMs (query)	0.408	0.541	0.813	0.802
LSTMs (context+query)	0.418	0.560	**0.818**	0.791

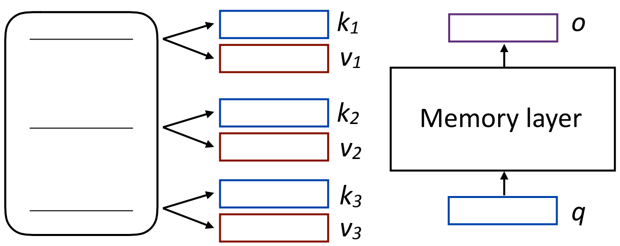
- Why are these results so low?

Hill et al. (2015)

Memory Networks

Memory Networks

- Memory networks let you reference input with attention
- Encode input items into two vectors: a key and a value
- Keys compute attention weights given a query, weighted sum of values gives the output



$$o = \sum_i \alpha_i v_i$$

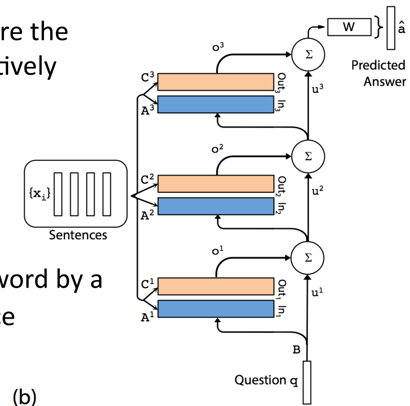
$$\alpha = \mathrm{softmax}(e)$$

$$e_i = q \cdot k_i$$

Sukhbaatar et al. (2015)

Memory Networks

- Three layers of memory network where the query representation is updated additively based on the memories at each step
- How to encode the sentences?
  - Bag of words (average embeddings)
  - Positional encoding: multiply each word by a vector capturing position in sentence



(b)

Sukhbaatar et al. (2015)

bAbI

- Evaluation on 20 tasks proposed as building blocks for building "AI-complete" systems
- Various levels of difficulty, exhibit different linguistic phenomena
- Small vocabulary, language isn't truly "natural"

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 13: Compound Coreference
Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning
In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Weston et al. (2014)

Evaluation: bAbI

	Baseline			MemN2N				
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint
Mean error (%)	6.7	51.3	40.2	25.1	20.3	25.8	15.6	13.3
Failed tasks (err. > 5%)	4	20	18	15	13	17	11	11

- 3-hop memory network does pretty well, better than LSTM at processing these types of examples

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			

Evaluation: Children's Book Test

METHODS	NAMED ENTITIES
HUMANS (QUERY)(*)	0.520
HUMANS (CONTEXT+QUERY)(*)	***0.816***
MAXIMUM FREQUENCY (CORPUS)	0.120
MAXIMUM FREQUENCY (CONTEXT)	0.335
SLIDING WINDOW	0.168
WORD DISTANCE MODEL	0.398
KNESER-NEY LANGUAGE MODEL	0.390
KNESER-NEY LANGUAGE MODEL + CACHE	0.439
LSTMS (QUERY)	0.408
LSTMS (CONTEXT+QUERY)	0.418
CONTEXTUAL LSTMS (WINDOW CONTEXT)	0.436
MEMNNS (LEXICAL MEMORY)	0.431
MEMNNS (WINDOW MEMORY)	0.493
MEMNNS (SENTENTIAL MEMORY + PE)	0.318
MEMNNS (WINDOW MEMORY + SELF-SUP.)	**0.666**

- Outperforms LSTMs substantially with the right supervision

Memory Network Takeaways

- Memory networks provide a way of attending to abstractions over the input
- Useful for cloze tasks where far-back context is necessary
- What can we do with more basic attention?

CNN/Daily Mail: Attentive Reader

CNN/Daily Mail

- Single-document, (usually) single-sentence cloze task
- Formed based on article summaries — information should mostly be present, makes it easier than Children's Book Test
- Need to process the question, can't just use LSTM LMs

Passage

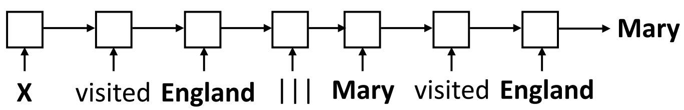
(@entity4) if you feel a ripple in the force today , it may be the news that the official @entity6 is getting its first gay character . according to the sci-fi website @entity9 , the upcoming novel " @entity11 " will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian . " the character is the first gay figure in the official @entity6 -- the movies , television shows , comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24 , editor of " @entity6 " books at @entity28 imprint @entity26 .

Question	Answer
characters in " @placeholder " movies have gradually become more diverse	@entity6

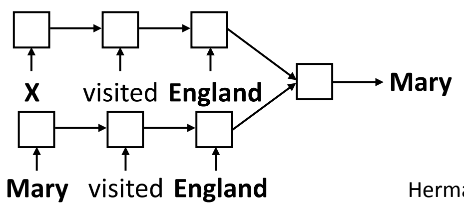
Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

- LSTM reader: encode question, encode passage, predict entity



- Can also use textual entailment-like models

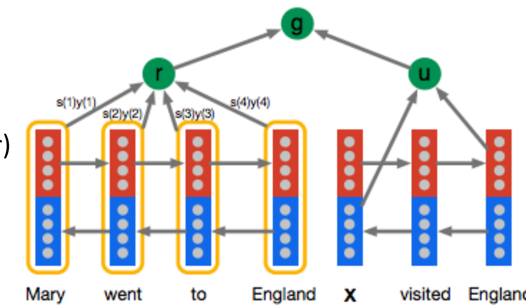


Multiclass classification problem over entities in the document

Hermann et al. (2015), Chen et al. (2016)

CNN/Daily Mail

- Attentive reader:
  u = encode query
  s = encode sentence
  r = attention(u -> s)
  prediction = f(candidate, u, r)
- Uses fixed-size representations for the final prediction, multiclass classification



Hermann et al. (2015)

CNN/Daily Mail

- Chen et al (2016): small changes to the attentive reader
- Additional analysis of the task found that many of the remaining questions were unanswerable or extremely difficult

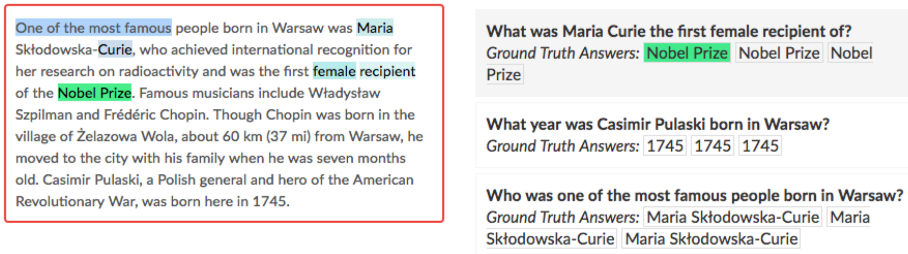
	CNN		Daily Mail	
	valid	test	valid	test
Maximum frequency	30.5	33.2	25.6	25.5
Exclusive frequency	36.6	39.3	32.7	32.8
Frame-semantic model	36.3	40.2	35.5	35.5
Word distance model	50.5	50.9	56.4	55.5
Deep LSTM Reader	55.0	57.0	63.3	62.2
Uniform Reader	39.0	39.4	34.6	34.4
Attentive Reader	61.6	63.0	**70.5**	**69.0**
Stanford Attentive Reader	76.2	76.5	**79.5**	**78.7**

Hermann et al. (2015), Chen et al. (2016)

SQuAD: Bidirectional Attention Flow

SQuAD

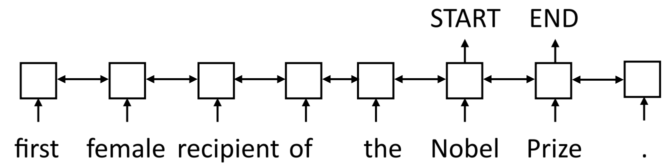
- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage



Rajpurkar et al. (2016)

SQuAD

What was Marie Curie the first female recipient of?



- Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query

Rajpurkar et al. (2016)

Bidirectional Attention Flow

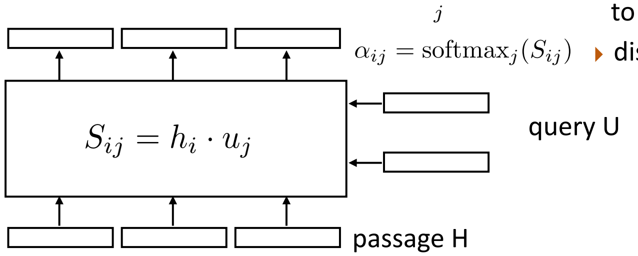
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

$$\tilde{u}_i = \sum_j \alpha_{ij} u_j$$

- query "specialized" to the *i*th word

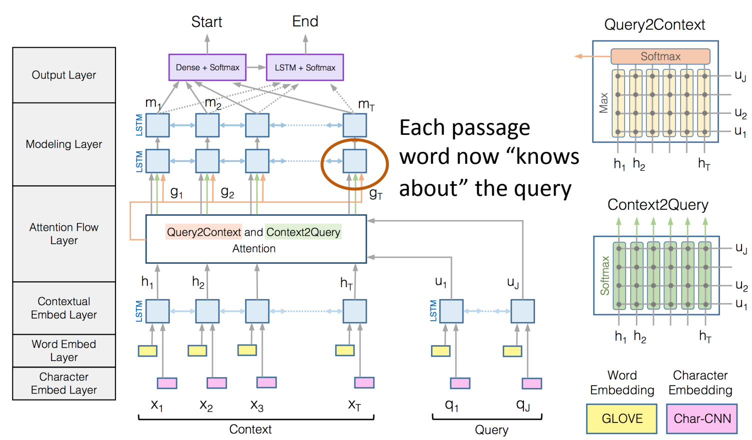
$$\alpha_{ij} = \text{softmax}_j(S_{ij})$$

- dist over query



Seo et al. (2016)

Bidirectional Attention Flow



Each passage word now "knows about" the query

Seo et al. (2016)

SQuAD SOTA

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF…)
- BERT: transformer-based approach with pretraining on 3B tokens

But how well are these doing?

- Can construct adversarial examples that fool these systems: add one carefully chosen sentence and performance drops to below 50%
- Still "surface-level" matching, not complex understanding
- Other challenges: recognizing when answers aren't present, doing multi-step reasoning

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Jia and Liang (2017)

Takeaways

- Many flavors of reading comprehension tasks: cloze or actual questions, single or multi-sentence
- Memory networks let you reference input in an attention-like way, useful for generalizing language models to long-range reasoning
- Complex attention schemes can match queries against input texts and identify answers