Representation Learning for Reading Comprehension

Russ Salakhutdinov

Machine Learning Department
Carnegie Mellon University
Canadian Institute for Advanced Research

Joint work with
Bhuwan Dhingra, Zhilin Yang, Ye Yuan, Junjie Hu,
Hanxiao Liu, and William Cohen

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Talk Roadmap

- Multiplicative and Fine-grained Attention
- Incorporating Knowledge as Explicit Memory for RNNs
- Generative Domain-Adaptive Nets

Who-Did-What Dataset

- **Document**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"

- **Query**: President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat.

- **Answer**: Rod Blagojevich

Onishi, Wang, Bansal, Gimpel, McAllester, EMNLP, 2016

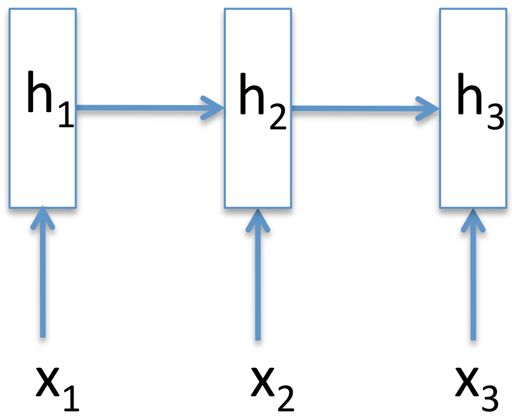
Recurrent Neural Network

$$\mathbf{h_t} = \phi(\mathbf{U}\mathbf{h_{t-1}} + \mathbf{W}\mathbf{x_t} + \mathbf{b})$$

Nonlinearity

Hidden State at previous time step

Input at time step t



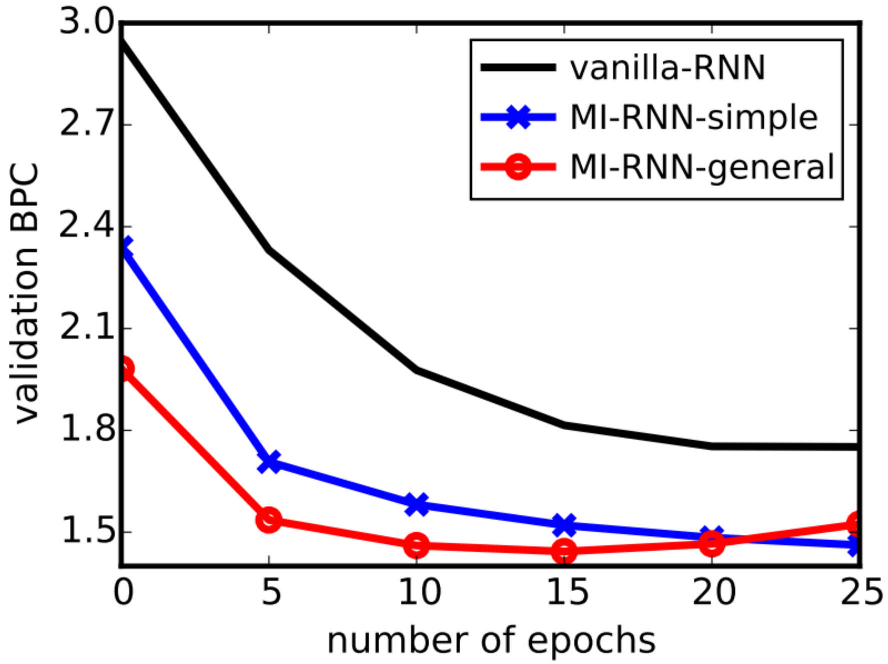
Multiplicative Integration

- Replace

$$\phi(\mathbf{U}\mathbf{h} + \mathbf{W}\mathbf{x} + \mathbf{b})$$

- With

$$\phi(\mathbf{U}\mathbf{h} \odot \mathbf{W}\mathbf{x} + \mathbf{b})$$

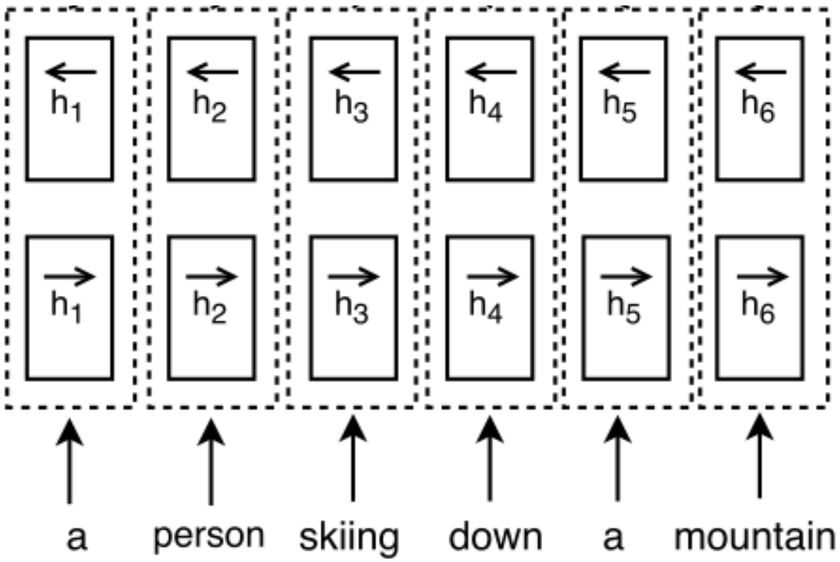


- Or more generally

$$\phi(\alpha \odot \mathbf{U}\mathbf{h} \odot \mathbf{W}\mathbf{x} + \beta_1 \odot \mathbf{U}\mathbf{h} + \beta_2 \odot \mathbf{W}\mathbf{x} + \mathbf{b})$$

Wu et al., NIPS 2016

Representing Document/Query



- Forward RNN reads sentences from left to right:

$$[\overrightarrow{h}_1, \overrightarrow{h}_2, .., \overrightarrow{h}_{|D|}]$$

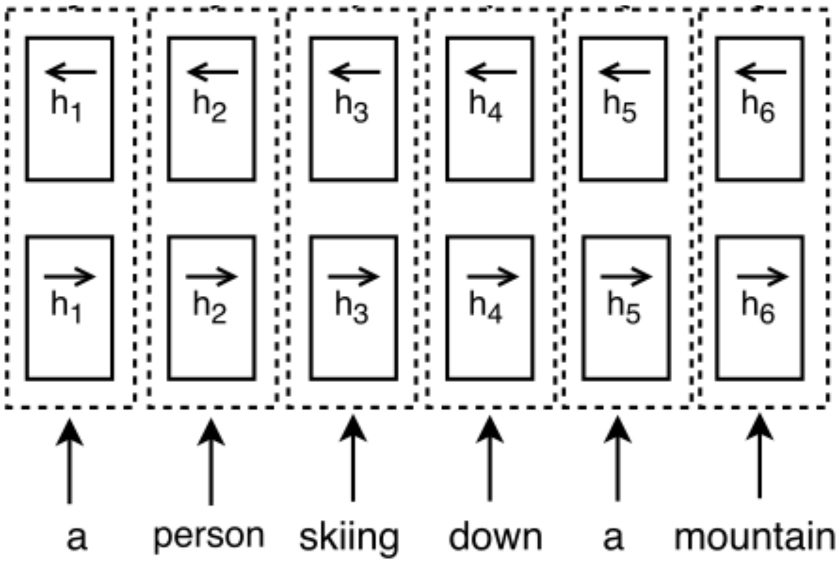
- Backward RNN reads sentences from right to left:

$$[\overleftarrow{h}_1, \overleftarrow{h}_2, .., \overleftarrow{h}_{|D|}]$$

- The hidden states are then concatenated:

$$\overleftrightarrow{\mathrm{GRU}} = [h_1, h_2, ..., h_{|D|}], \quad h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i]$$

Representing Document/Query



- Use GRUs to encode a document and a query:

$$D = \overleftrightarrow{\mathrm{GRU}}_D(X)$$

$$Q = \overleftrightarrow{\mathrm{GRU}}_Q(Y)$$

- Note that, for example, Q is a matrix

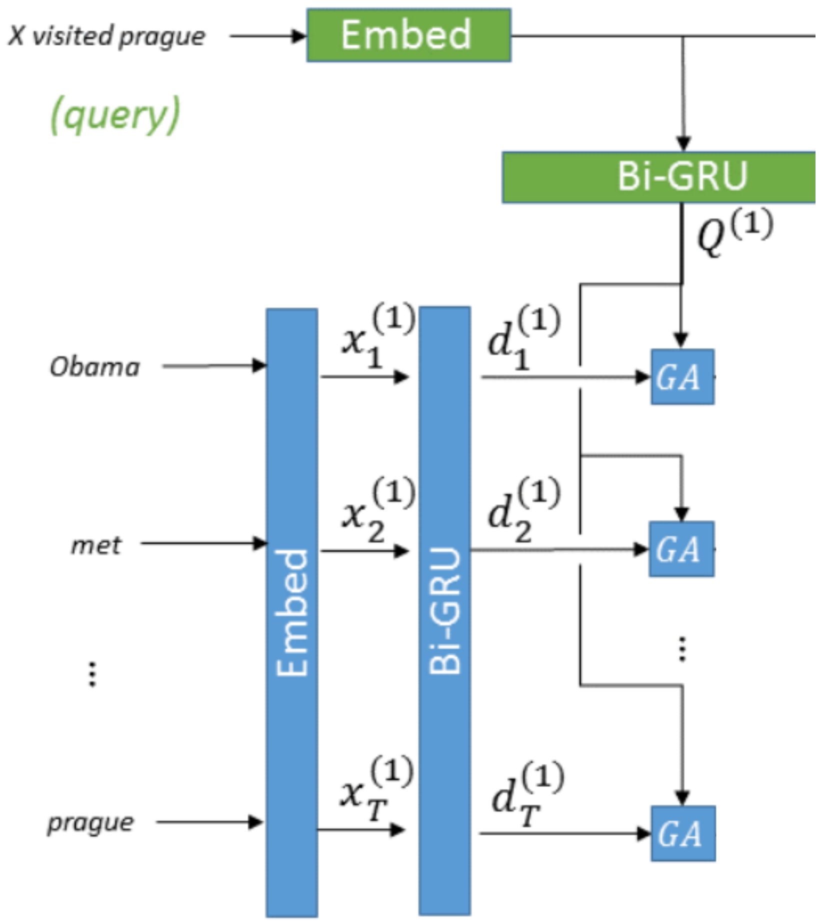
$$Q \in \mathbb{R}^{2|H| \times |Q|}$$

- We can then use Gated Attention mechanism:

$$X = \mathrm{GA}(D, Q)$$

Gated Attention Mechanism

- For each token d in D, we form a token-specific representation of the query:



$$\alpha_i = \mathrm{softmax}(Q^\top d_i)$$

$$\tilde{q}_i = Q\alpha_i$$

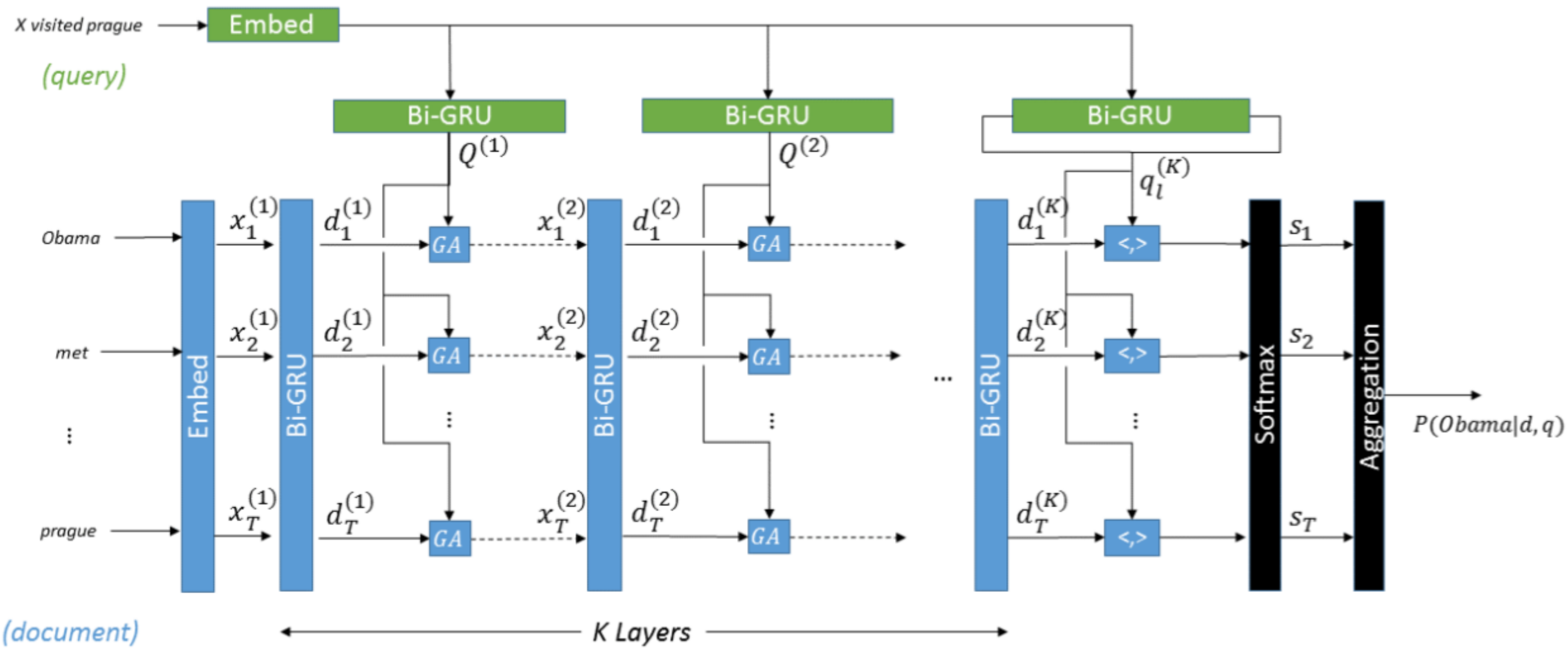
$$\boxed{x_i = d_i \odot \tilde{q}_i}$$

➢ use the element-wise multiplication operator to model the interactions between d_i and $\tilde{q}_i$

Dhingra, Liu, Yang, Cohen, Salakhutdinov, ACL 2017

Multi-hop Architecture

- Many QA tasks require reasoning over multiple sentences.
- Need to performs several passes over the context.



Dhingra, Liu, Yang, Cohen, Salakhutdinov, ACL 2017

Affect of Multiplicative Gating

- Performance of different gating functions on "Who did What" (WDW) dataset.

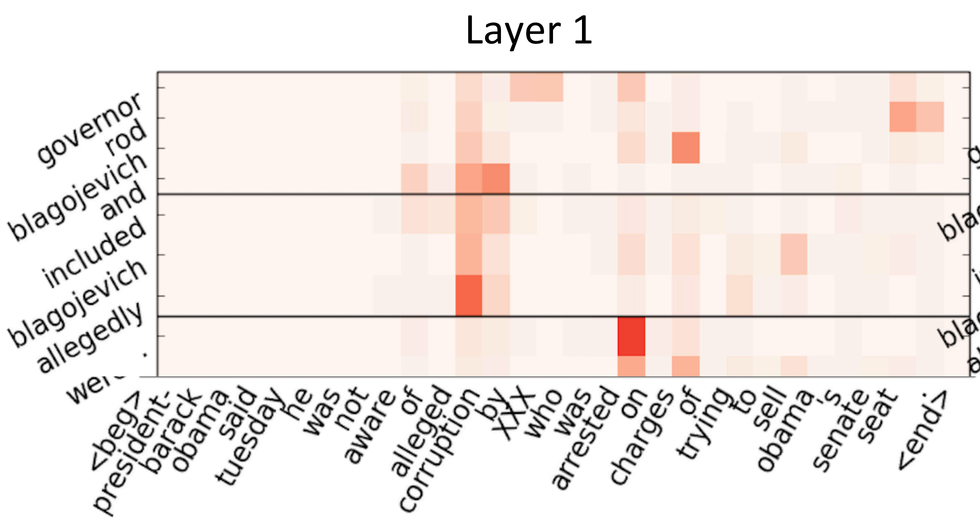
Gating Function	Accuracy	
	Val	Test
Sum	64.9	64.5
Concatenate	64.4	63.7
Multiply	**68.3**	**68.0**

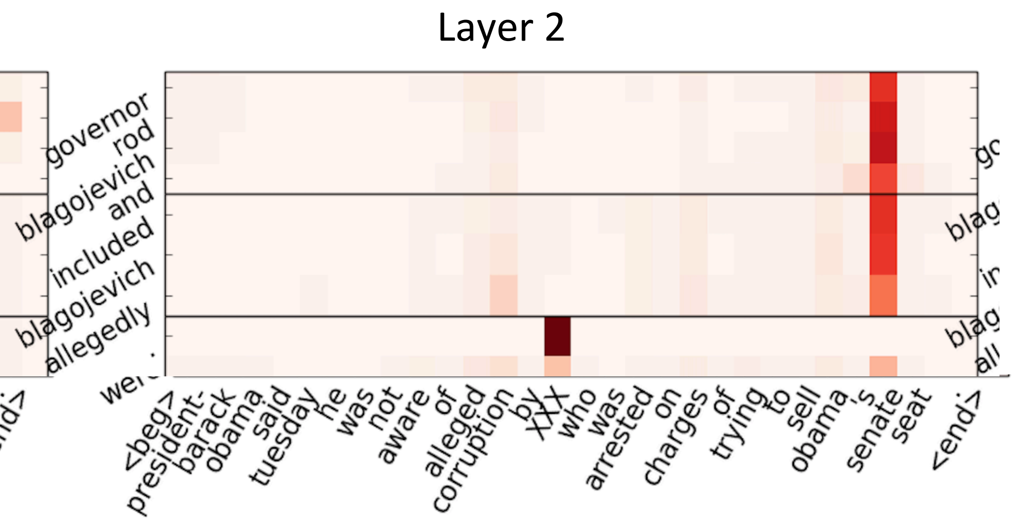
Model	Strict		Relaxed	
	Val	Test	Val	Test
Human †	–	84.0	–	–
Attentive Reader †	–	53.0	–	55.0
AS Reader †	–	57.0	–	59.0
Stanford AR †	–	64.0	–	65.0
NSE †	66.5	66.2	67.0	66.7
GA-- †	–	57.0	–	60.0
GA (update $L(w)$)	67.8	67.0	67.0	66.6
GA (fix $L(w)$)	68.3	68.0	69.6	69.1
GA (+feature, update $L(w)$)	70.1	69.5	70.9	71.0
GA (+feature, fix $L(w)$)	**71.6**	**71.2**	**72.6**	**72.6**

Model	CNN		Daily Mail		CBT-NE		CBT-CN	
	Val	Test	Val	Test	Val	Test	Val	Test
Humans (query) †	–	–	–	–	–	52.0	–	64.4
Humans (context + query) †	–	–	–	–	–	81.6	–	81.6
LSTMs (context + query) †	–	–	–	–	51.2	41.8	62.6	56.0
Deep LSTM Reader †	55.0	57.0	63.3	62.2	–	–	–	–
Attentive Reader †	61.6	63.0	70.5	69.0	–	–	–	–
Impatient Reader †	61.8	63.8	69.0	68.0	–	–	–	–
MemNets †	63.4	66.8	–	–	70.4	66.6	64.2	63.0
AS Reader †	68.6	69.5	75.0	73.9	73.8	68.6	68.8	63.4
DER Network †	71.3	72.9	–	–	–	–	–	–
Stanford AR (relabeling) †	73.8	73.6	77.6	76.6	–	–	–	–
Iterative Attentive Reader †	72.6	73.3	–	–	75.2	68.6	72.1	69.2
EpiReader †	73.4	74.0	–	–	75.3	69.7	71.5	67.4
AoA Reader †	73.1	74.4	–	–	77.8	72.0	72.2	69.4
ReasoNet †	72.9	74.7	77.6	76.6	–	–	–	–
NSE †	–	–	–	–	78.2	73.2	74.3	**71.9**
MemNets (ensemble) †	66.2	69.4	–	–	–	–	–	–
AS Reader (ensemble) †	73.9	75.4	78.7	77.7	76.2	71.0	71.1	68.9
Stanford AR (relabeling,ensemble) †	77.2	77.6	80.2	79.2	–	–	–	–
Iterative Attentive Reader (ensemble) †	75.2	76.1	–	–	76.9	72.0	74.1	71.0
EpiReader (ensemble) †	–	–	–	–	76.6	71.8	73.6	70.6
AS Reader (+BookTest) † ‡	–	–	–	–	80.5	76.2	83.2	80.8
AS Reader (+BookTest,ensemble) † ‡	–	–	–	–	*82.3*	*78.4*	*85.7*	*83.7*
GA--	73.0	73.8	76.7	75.7	74.9	69.0	69.0	63.9
GA (update $L(w)$)	**77.9**	**77.9**	**81.5**	**80.9**	76.7	70.1	69.8	67.3
GA (fix $L(w)$)	77.9	77.8	80.4	79.6	77.2	71.4	71.6	68.0
GA Reader (+feature, update $L(w)$)	77.3	76.9	80.7	80.0	77.2	73.3	73.0	69.8
GA Reader (+feature, fix $L(w)$)	76.7	77.4	80.0	79.3	**78.5**	**74.9**	**74.4**	70.7

Analysis of Attention

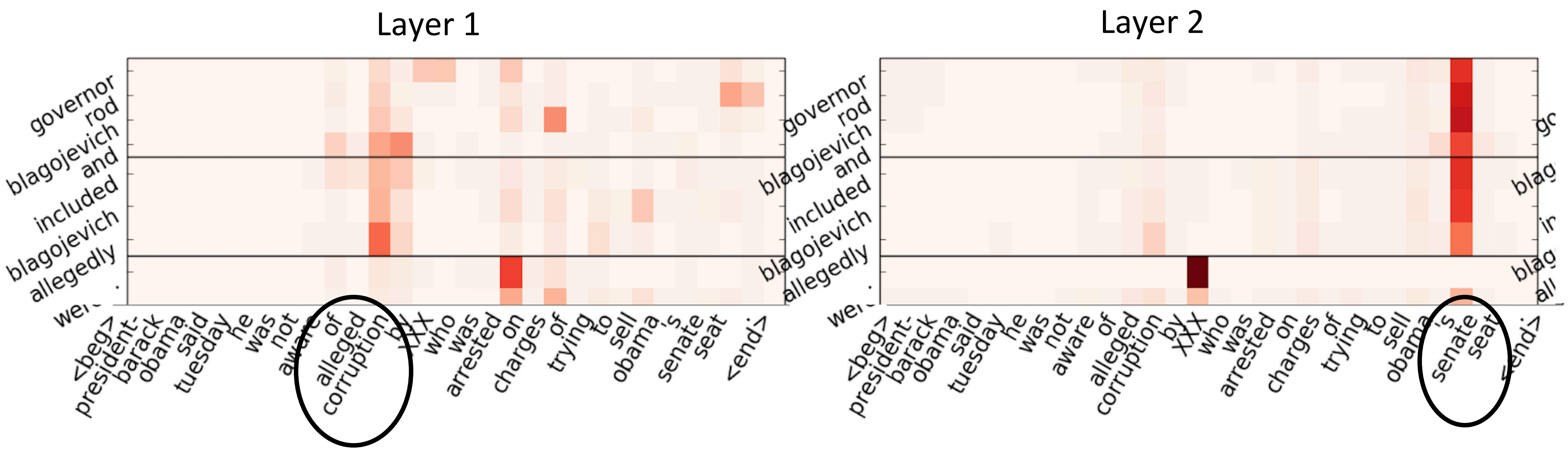
- **Context**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"
- **Query**: "President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
- **Answer**: **Rod Blagojevich**





Analysis of Attention

- **Context**: "…arrested Illinois governor **Rod Blagojevich** and his chief of staff John Harris on corruption charges … included **Blogojevich** allegedly conspiring to sell or trade the senate seat left vacant by President-elect Barack Obama…"
- **Query**: "President-elect Barack Obama said Tuesday he was not aware of alleged corruption by **X** who was arrested on charges of trying to sell Obama's senate seat."
- **Answer**: **Rod Blagojevich**



Code + Data: https://github.com/bdhingra/ga-reader

Words vs. Characters

- Word-level representations are good at learning the semantics of the tokens

- Character-level representations are more suitable for modeling sub-word morphologies (“cat” vs. “cats”)

- Hybrid word-character models have been shown to be successful in various NLP tasks (Yang et al., 2016a, Miyamoto & Cho (2016), Ling et al., 2015)

Fine-Grained Gating

- Fine-grained gating mechanism:

$$\mathbf{h} = \mathbf{g} \odot \mathbf{c} + (1 - \mathbf{g}) \odot (\mathbf{E}\mathbf{w})$$

Character - level representation

Gating

Word- level representation

$$\mathbf{g} = \sigma(\mathbf{W}_g \mathbf{v} + \mathbf{b}_g)$$

Additional features: named entity tags, part- of-speech tags, document frequency vectors, word look-up representations

Yang et al, ICLR 2017

Children’s Book Test (CBC) Dataset

Model	CN dev	CN test	NE dev	NE test
GA word char concat	0.731	0.696	0.768	0.725
GA word char feat concat	0.7250	0.6928	0.7815	0.7256
GA scalar gate	0.7240	0.6908	0.7810	0.7260
GA fine-grained gate	0.7425	0.7084	0.7890	0.7464
FG fine-grained gate	**0.7530**	**0.7204**	**0.7910**	**0.7496**
Sordoni et al. (2016)	0.721	0.692	0.752	0.686
Trischler et al. (2016)	0.715	0.674	0.753	0.697
Cui et al. (2016)	0.722	0.694	0.778	0.720
Munkhdalai & Yu (2016)	0.743	0.719	0.782	0.732
Kadlec et al. (2016) ensemble	0.711	0.689	0.762	0.710
Sordoni et al. (2016) ensemble	0.741	0.710	0.769	0.720
Trischler et al. (2016) ensemble	0.736	0.706	0.766	0.718

Words vs. Characters

- High gate values: character-level representations
- Low gate values: word-level representations.

Gate values	Word tokens
Lowest	or but But These these However however among Among that when When although Although because Because until many Many than though Though this This Since since date where Where have That and And Such such number so which by By how before Before with With between Between even Even if
Highest	Sweetgum Untersee Jianlong Floresta Chlorella Obersee PhT Doctorin Jumonville WFTS WTSP Boven Pharm Nederrijn Otrar Rhin Magicicada WBKB Tanzler KMBC WPLG Mainau Merwede RMJM Kleitman Scheur Bodensee Kromme Horenbout Vorderrhein Chlamydomonas Scantlebury Qingshui Funchess

Talk Roadmap

- Multiplicative and Fine-grained Attention
- Linguistic Knowledge as Explicit Memory for RNNs
- Generative Domain-Adaptive Nets

Broad-Context Language Modeling

Her plain face broke into a huge smile when she saw Terry.
“Terry!” she called out.
She rushed to meet him and they embraced.
“Hon, I want you to meet an old friend, Owen McKenna.
Owen, please meet Emily."
She gave me a quick nod and turned back to **X**

LAMBADA dataset, Paperno et al., 2016

Broad-Context Language Modeling

Her plain face broke into a huge smile when she saw **Terry**.
“Terry!” she called out.
She rushed to meet him and they embraced.
“Hon, I want you to meet an old friend, **Owen** McKenna.
Owen, please meet **Emily**."
She gave me a quick nod and turned back to **X**

LAMBADA dataset, Paperno et al., 2016

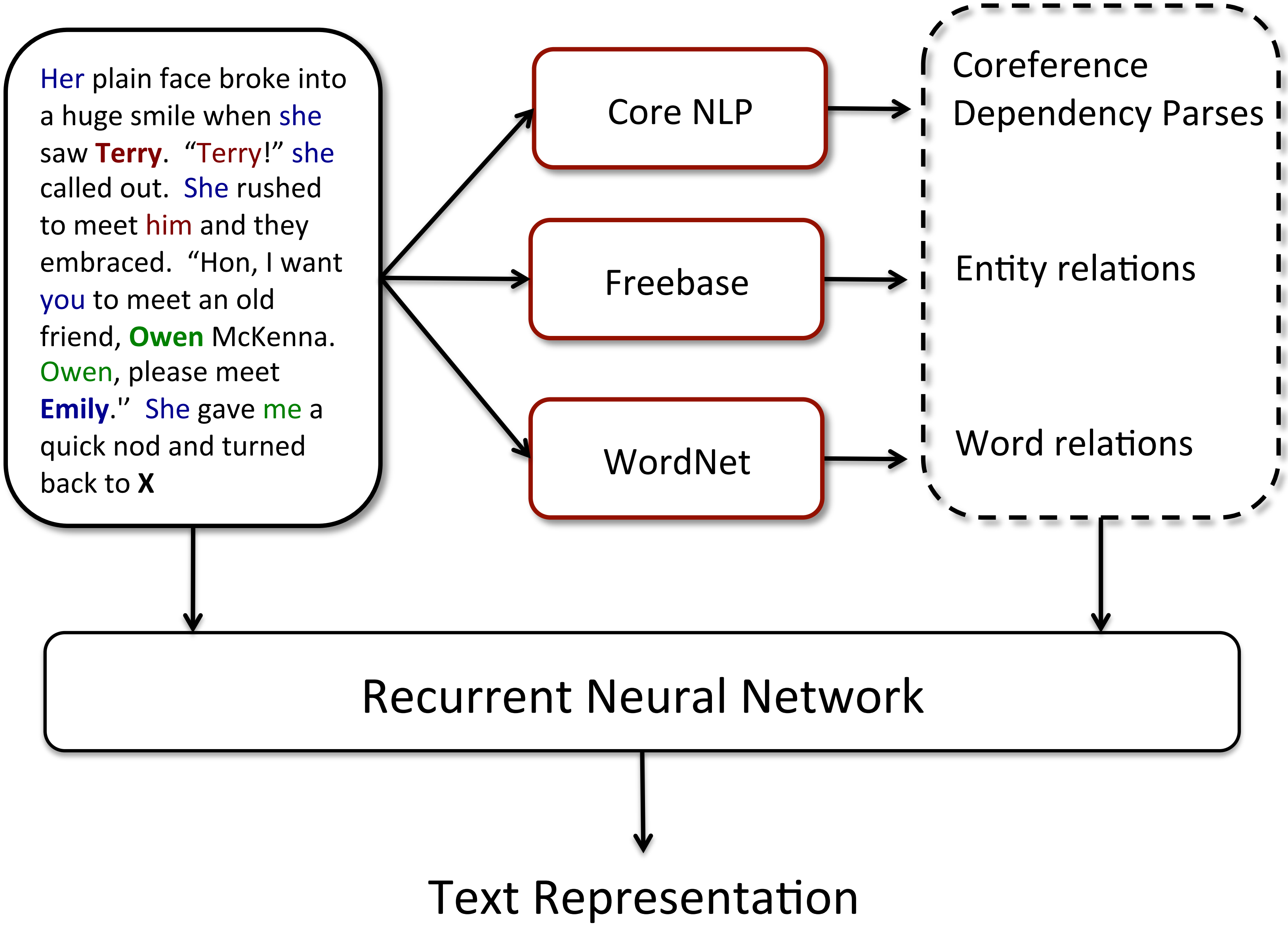
Broad-Context Language Modeling

Her plain face broke into a huge smile when she saw **Terry**.
"Terry!" she called out.
She rushed to meet him and they embraced.
"Hon, I want you to meet an old friend, **Owen** McKenna.
Owen, please meet **Emily**."
She gave me a quick nod and turned back to **X**

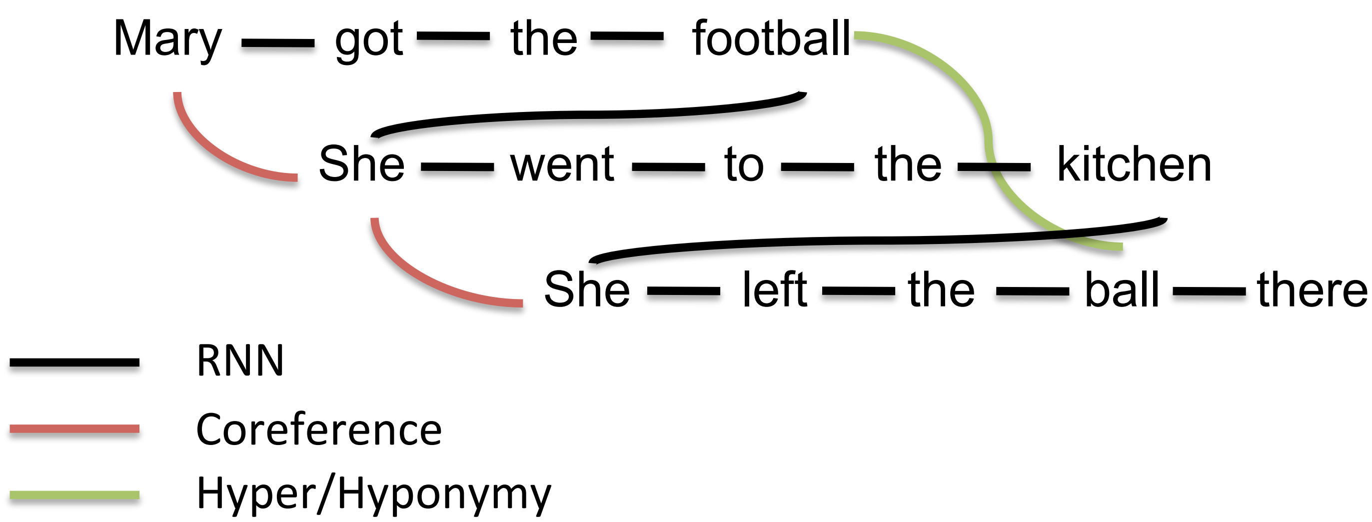
X = **Terry**

LAMBADA dataset, Paperno et al., 2016

Incorporating Prior Knowledge

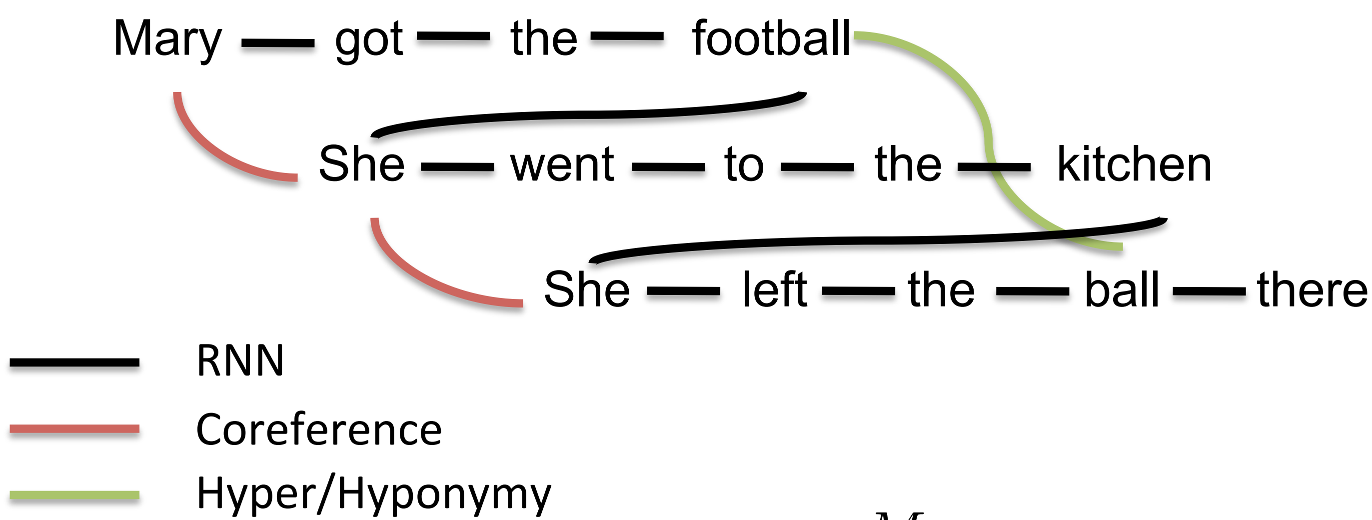


Incorporating Prior Knowledge

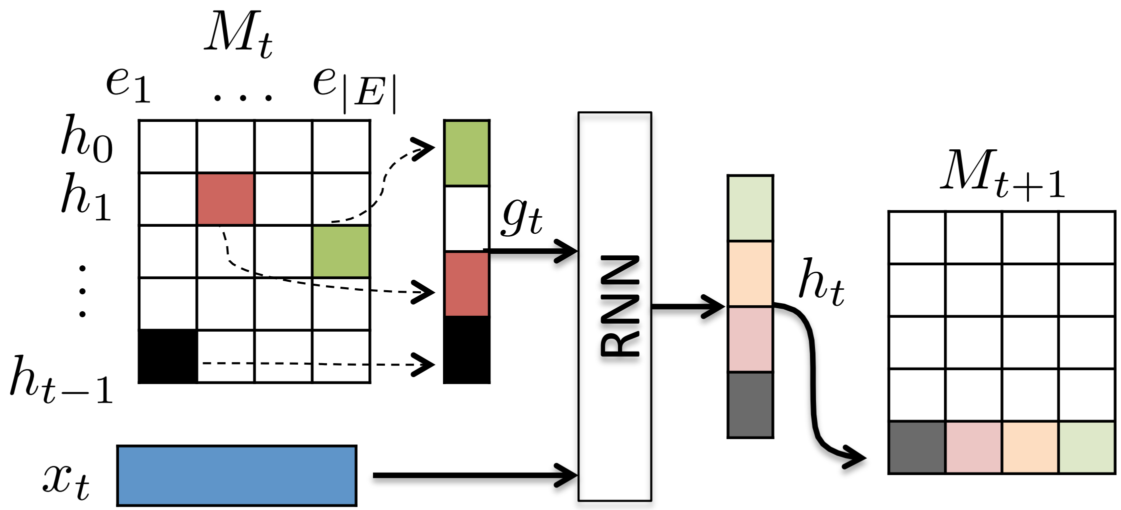


Dhingra, Yang, Cohen, Salakhutdinov 2017

Incorporating Prior Knowledge

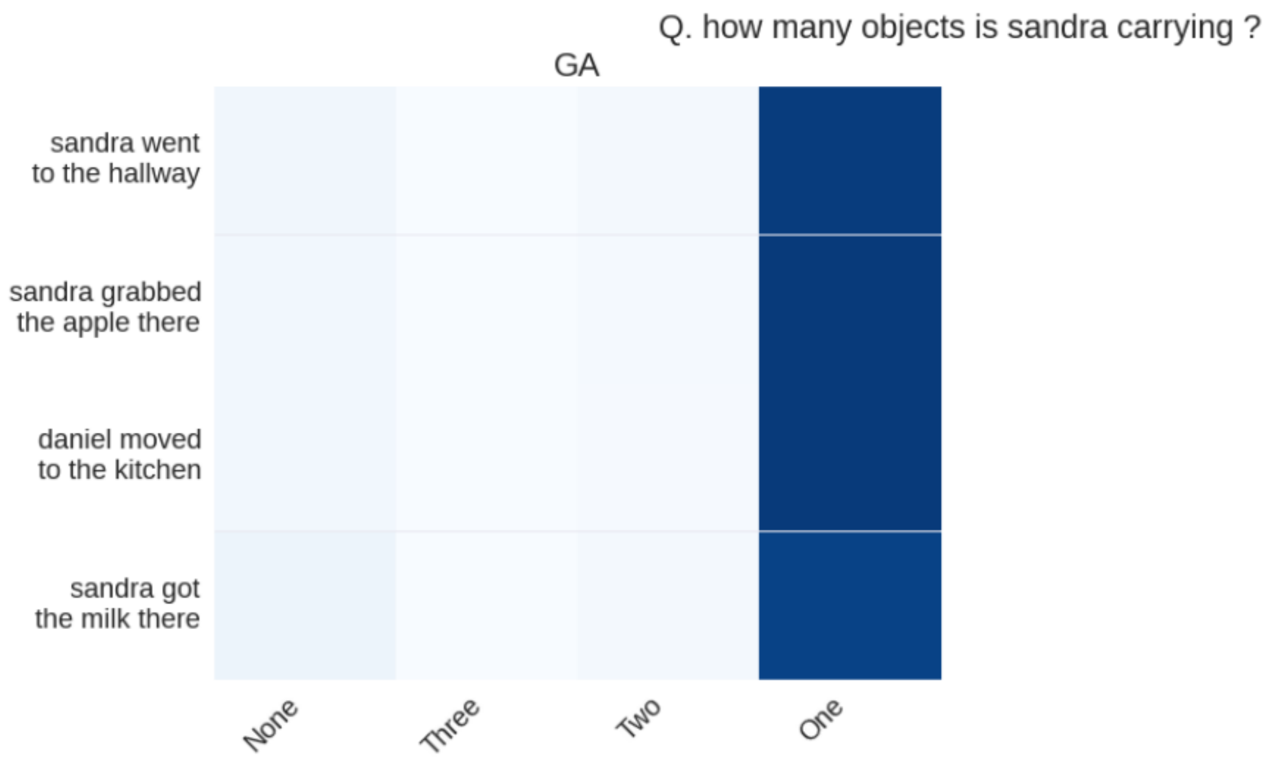


Memory as **A**cyclic **G**raph **E**ncoding (MAGE) - RNN

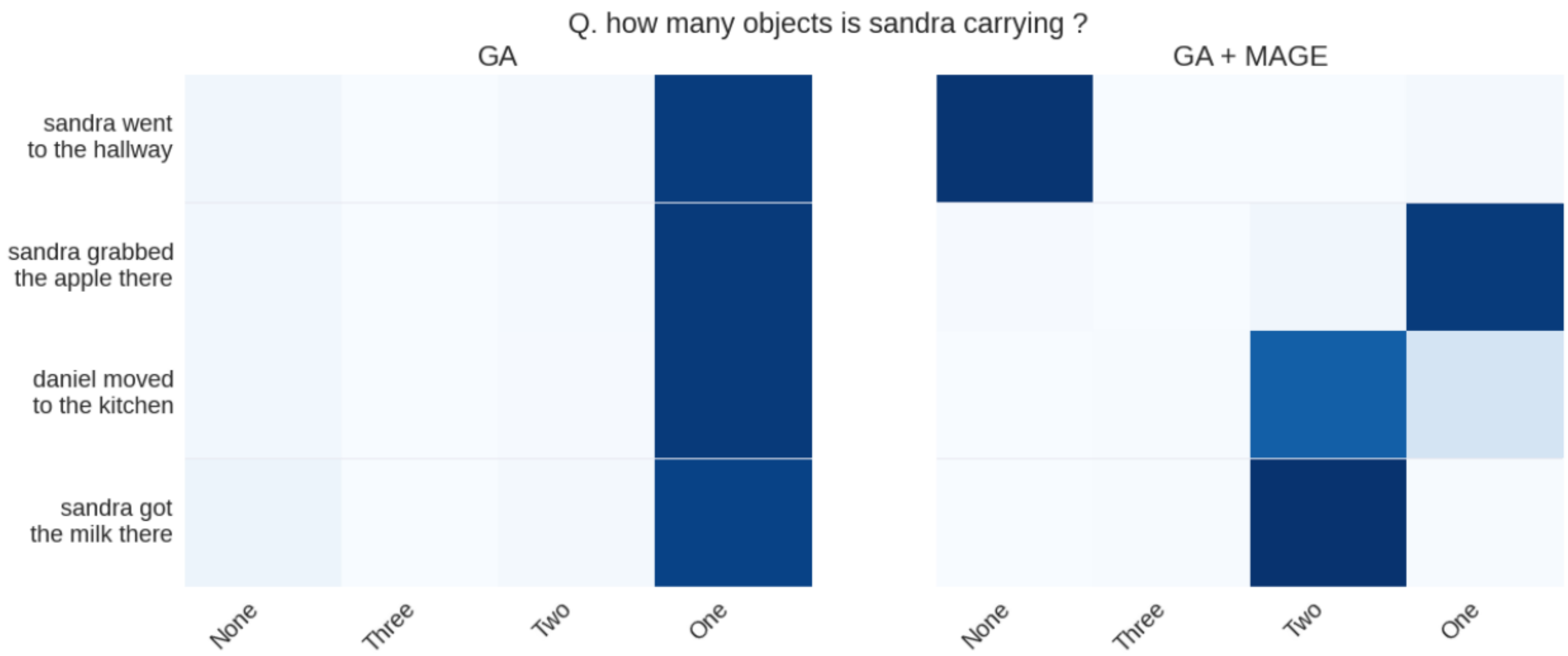


Dhingra, Yang, Cohen, Salakhutdinov 2017

Learned Representation



Learned Representation



Talk Roadmap

- Multiplicative and Fine-grained Attention
- Linguistic Knowledge as Explicit Memory for RNNs
- Generative Domain-Adaptive Nets

Extractive Question Answering

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, and hail… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers"

What causes precipitation to fall?
gravity

- Given a paragraph/question, extract a span of text as the answer
- Expensive to obtain large labeled datasets
- SOTA approaches rely on large labeled datasets

SQuAD Dataset, Rajpurkar et al., 2016

Leverage Unlabeled Text

Pittsburgh Steelers

From Wikipedia, the free encyclopedia
(Redirected from Steelers)

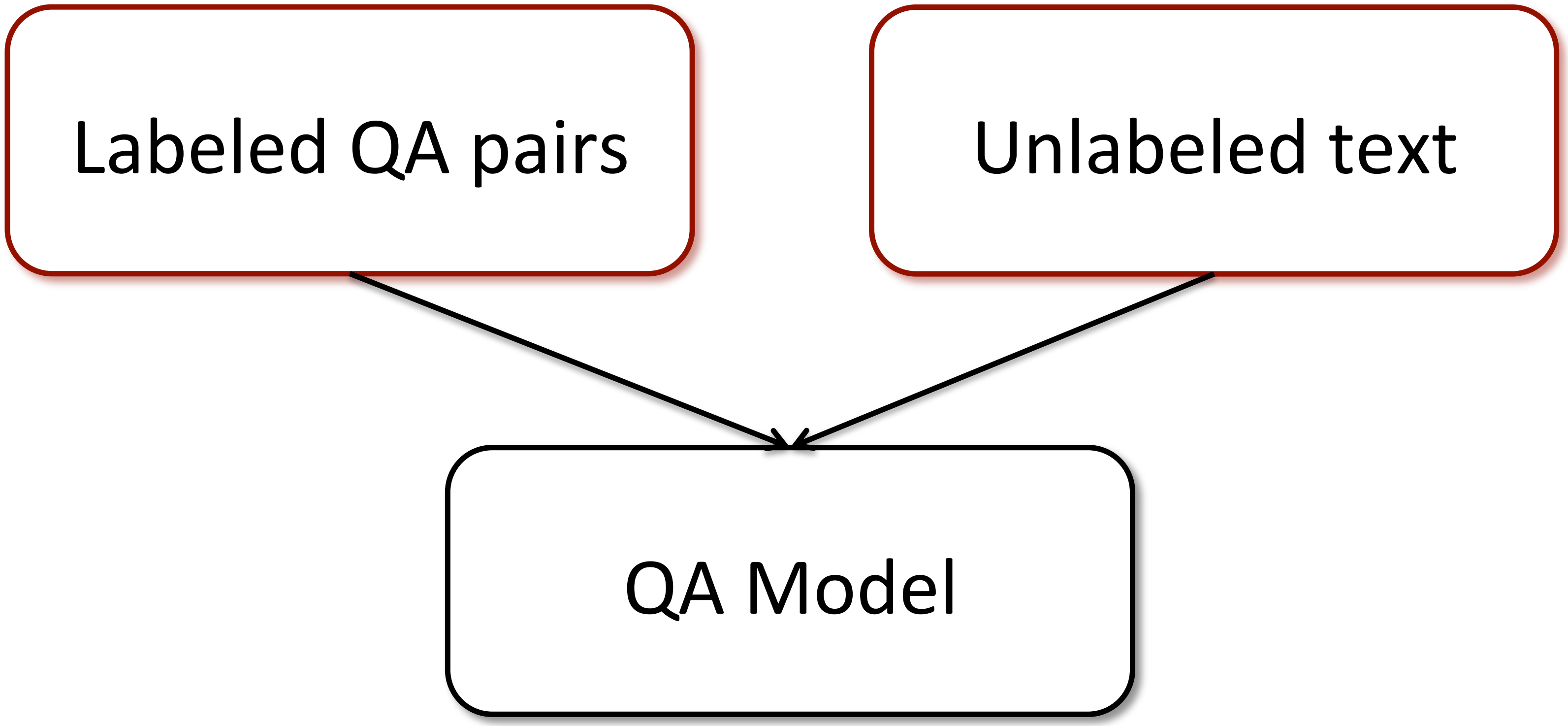
"Steelers" redirects here. For other uses, see Steelers (disambiguation).

The **Pittsburgh Steelers** are a professional American football team based in Pittsburgh, Pennsylvania.
Conference (AFC) North division. Founded in 1933, the Steelers are the oldest franchise in the AFC.

In contrast with their status as perennial also-rans in the pre-merger NFL, where they were the oldest te
successful NFL franchises. Pittsburgh has won more Super Bowl titles (6) and hosted more conference
Denver Broncos, but behind the New England Patriots record 9 AFC championships. They share the rec
record for second most Super Bowl appearances with the Broncos, and Dallas Cowboys (8), but again b

- Almost unlimited unlabeled text.

Semi-Supervised QA



Extractive Question Answering

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, and hail… Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers"

What causes precipitation to fall?
gravity

- Use POS/NER/parsing to extract possible answer chunks
- Anything can be the answers
- We will assume that answers are available.

Generating Questions

Labeled data

p,q,a

Unlabeled data

p, a

q

Generator G:

From (p, a) → q

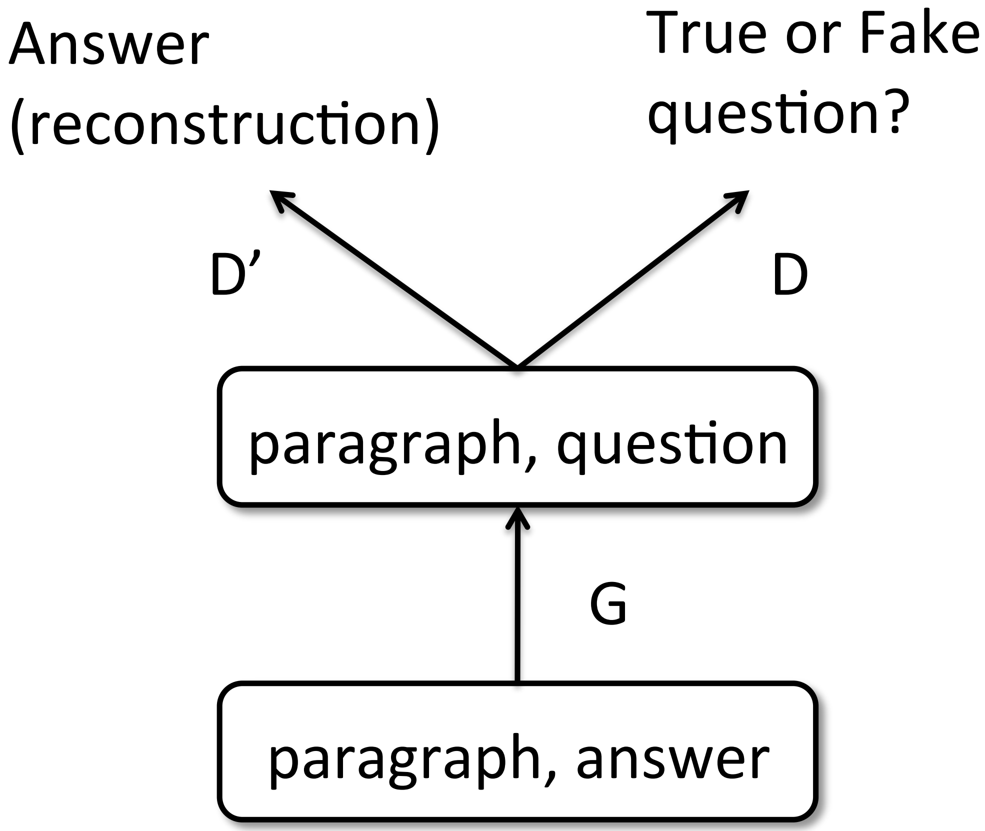
Seq2seq with copy mechanism

Discriminator D: Combine to train a QA model

From (p, q) → a

GA reader

Baseline: GANs



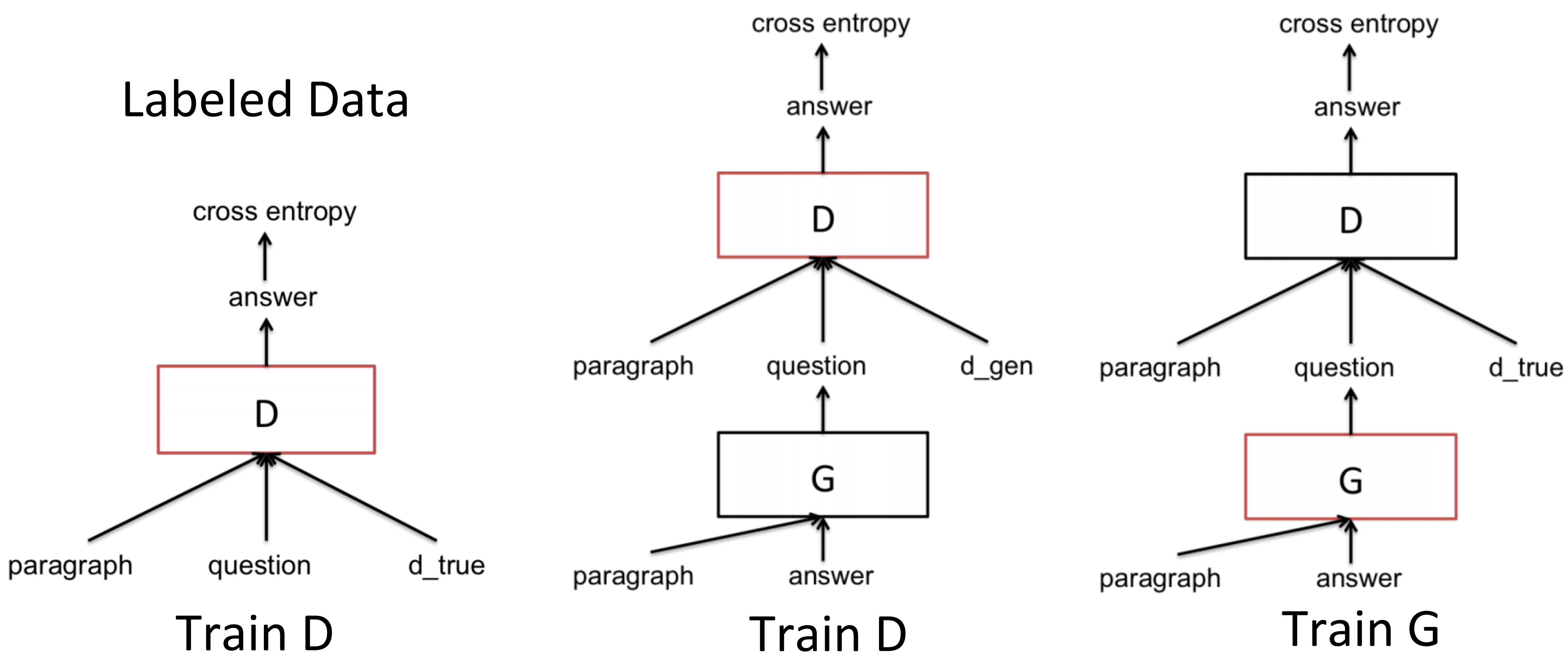
Goodfellow et al., 2014, Ganin et al. 2014 , Xia et al., 2016

Generative Domain-Adaptive Nets (GDANs)

$$\max_D \mathbb{E}_{data} \log p_D(y|x, \text{d\_true}) + \mathbb{E}_G \log p_D(y|x, \underline{\text{d\_gen}})$$

$$\max_G \mathbb{E}_G \log p_D(y|x, \underline{\text{d\_true}})$$

Unlabeled Data



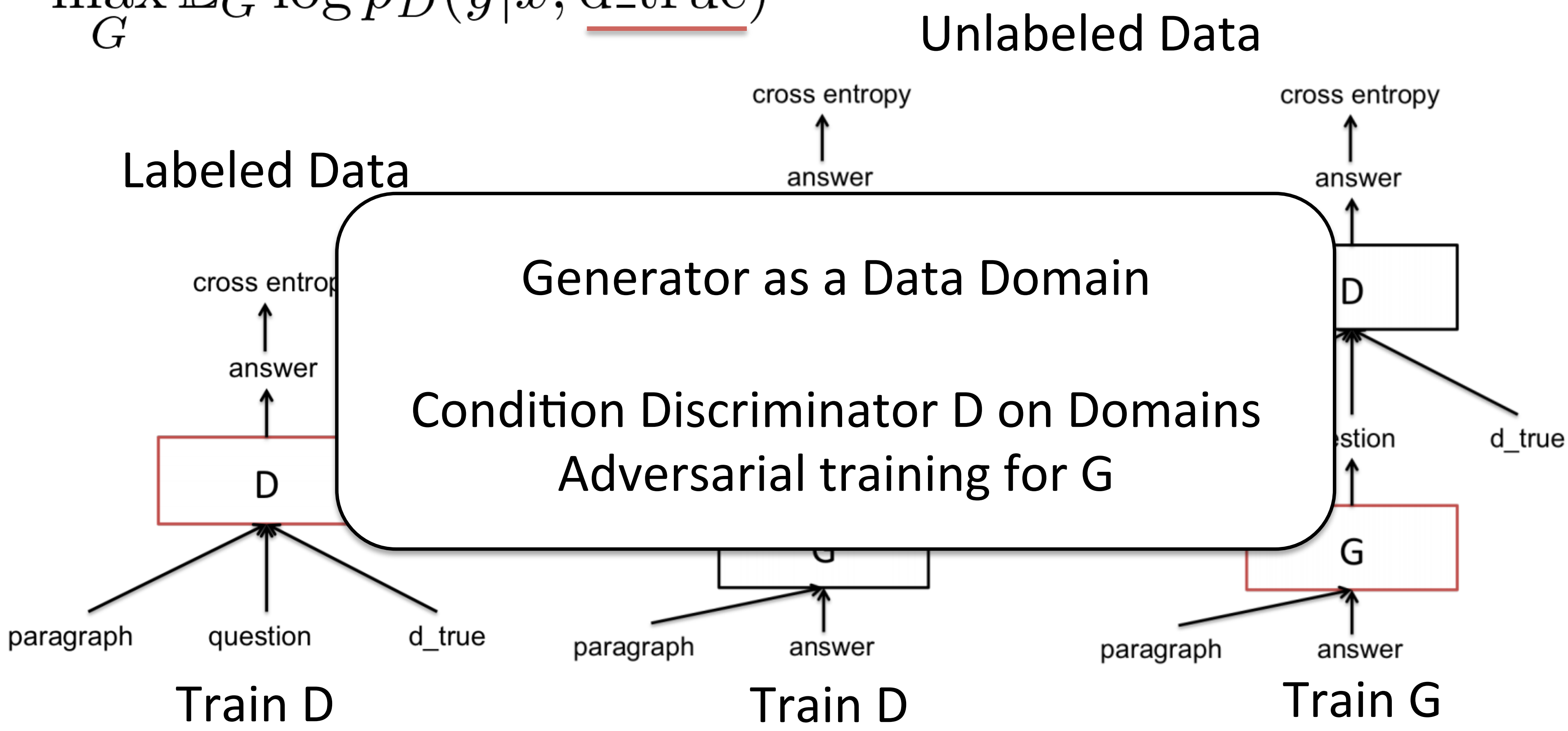
Johnson et al., 2016; Chu et al., 2017

Yang Hu Salakhutdinov, Cohen., ACL 2017

Generative Domain-Adaptive Nets (GDANs)

$$\max_D \mathbb{E}_{data} \log p_D(y|x, \text{d\_true}) + \mathbb{E}_G \log p_D(y|x, \underline{\text{d\_gen}})$$

$$\max_G \mathbb{E}_G \log p_D(y|x, \underline{\text{d\_true}})$$



Generator as a Data Domain

Condition Discriminator D on Domains

Adversarial training for G

Johnson et al., 2016; Chu et al., 2017

Yang Hu Salakhutdinov, Cohen., ACL 2017

Examples

Context: "…an additional warming of the Earth's surface. They calculate with confidence that C02 has been responsible for over half the enhanced greenhouse effect. They predict that under a "business as usual" scenario,…"

Answer: over half

Question: what the enhanced greenhouse effect that CO2 been responsible for?

Ground True Q: How much of the greenhouse effect is due to carbon dioxide?

Context: "… in 0000 , bankamericard was renamed and spun off into a separate company known today as visa inc."

Answer: visa inc .

Question: what was the separate company bankamericard?

Ground True Q: what present-day company did bankamericard turn into?

SQuAD dataset

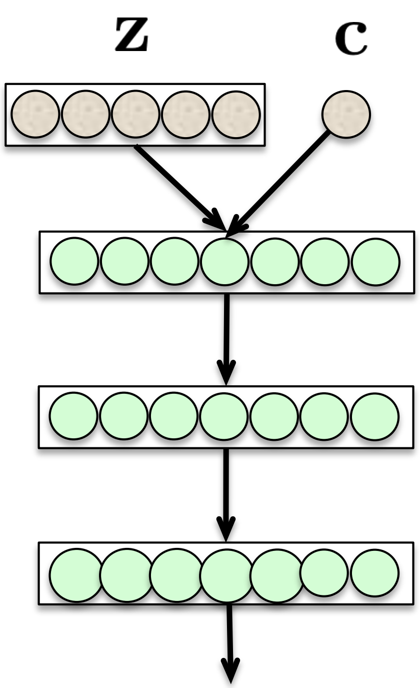
- SQuAD dataset: 87,636 training, 10,600 development instances
- Use 50K unlabelled examples.

Labeling rate	Method	Test F1	Exact Matching
0.1	Supervised	0.3815	0.2492
0.1	Context	0.4515	0.2966
0.1	Gen + GAN	0.4373	0.2885
0.1	GDAN	**0.4802**	**0.3218**
0.5	Supervised	0.5722	0.4187
0.5	Context	0.5740	0.4195
0.5	Gen + GAN	0.5590	0.4044
0.5	GDAN	**0.5831**	**0.4267**

Variational Autoencoder (VAE)

- Transform samples from some simple distribution (e.g. normal) to the data manifold:

$$p(\mathbf{x}, \mathbf{z}, \mathbf{c}) = p(\mathbf{x}|\mathbf{z}, \mathbf{c})p(\mathbf{z})p(\mathbf{c})$$



Generative Process

Deterministic neural network

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbf{c} \sim \text{Bernoulli}(\boldsymbol{\mu})$$

The movie was awful and boring

Knigma and Welling, 2014

VAE for Text Generation

- Sample c, fix z.

Varying the code of sentiment	Varying the code of tense
this movie was awful and boring . this movie was funny and touching .	this was one of the outstanding thrillers of the last decade this is one of the outstanding thrillers of the all time this will be one of the great thrillers of the all time
jackson is n't very good with documentary jackson is superb as a documentary productions	i thought the movie was too bland and too much i guess the movie is too bland and too much i guess the film will have been too bland
you will regret it you will enjoy it	

Hu, Yang, Liang, Salakhutdinov, Xing, ICML 2017

VAE for Text Generation

- Sample z, fix c.

Varying the unstructured code z	
("negative", "past")	*("positive", "past")*
the acting was also kind of hit or miss .	his acting was impeccable
i wish i 'd never seen it	this was spectacular , i saw it in theaters twice
by the end i was so lost i just did n't care anymore	it was a lot of fun
("negative", "present")	*("positive", "present")*
the movie is very close to the show in plot and characters	this is one of the better dance films
the era seems impossibly distant	i 've always been a big fan of the smart dialogue .
i think by the end of the film , it has confused itself	i recommend you go see this, especially if you hurt
("negative", "future")	*("positive", "future")*
i wo n't watch the movie	i hope he 'll make more movies in the future
and that would be devastating !	i will definitely be buying this on dvd
i wo n't get into the story because there really is n't one	you will be thinking about it afterwards, i promise you

Hu, Yang, Liang, Salakhutdinov, Xing, ICML 2017

Thank you