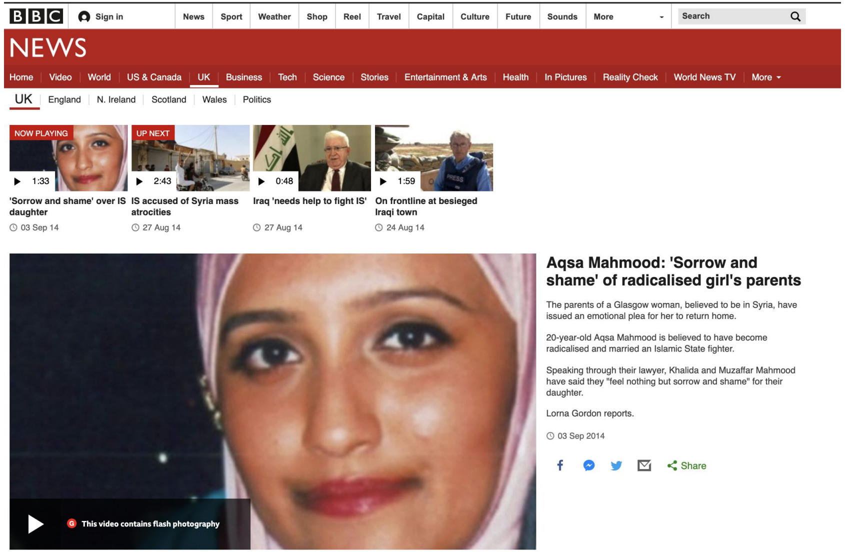
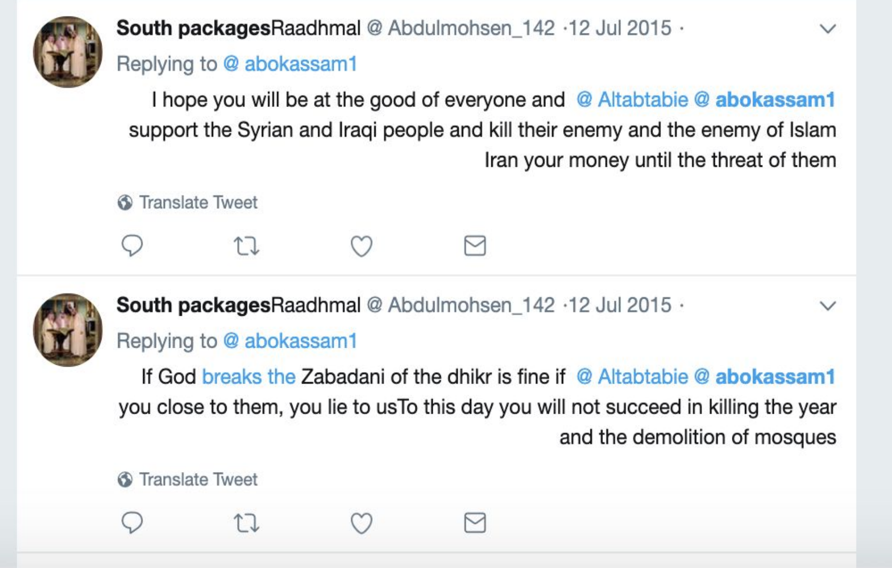
ISIL's Gender-Oriented Targeting on Twitter

- Who is ISIL?
- ISIL-gender-related headlines
- Data: 18-million tweets from ISIL-related accounts





A (relatively benign) ISIL-related Twitter account

ISIL Tweet Analysis Approach

- Genderize tweets by target
  - Evidence of target: author, retweet & mention
- Tag random words by the gender targeted
  - Generate word embeddings with tagged corpus
- Map word embeddings into a space where a specific dimension represents gender
  - Use stochastic gradient descent to push together same gender and pull apart separate genders for the gender dimension
  - Orthogonal mapping → preserve distances between embeddings

“...Mujahideen brothers in the state of Salah al-Din and everywhere O Allah support them against their enemies and accept their martyrs...”

Example **female-targeting** tweet

“...continued not to spend on the Messenger of Allah peace be upon him...”

Example **male-targeting** tweet

Learning from ISIL's Tweets

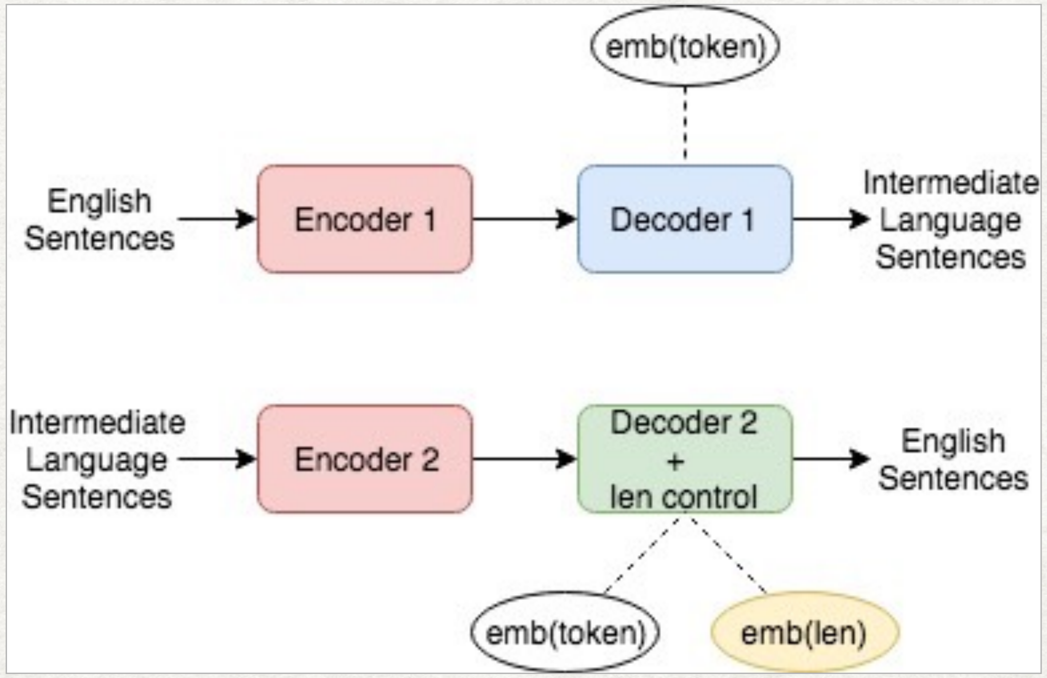
- Find a linear transformation: Map from <word>_m to <word>_f
  - $R^2 = 0.73$
- Densified embedding → analyze gender and non-gender dimensions

Notable masculine words	**Notable feminine words**	**Notable similarities (non-gender dimensions)**
rulers Mossad (Israeli intelligence) urgent militias industry puppet clients	Hayat ("life" or a city) injured replace careful	resurrection vs. judgment injuring vs. wounding Khafafa (female name) vs. Anfroa (cafe/bar) haha vs. hahaha

COMPARISON OF PIVOT LANGUAGES FOR AUTOMATIC SENTENCE COMPRESSION

Chaitra Hegde, Vish Rajiv, Ben Stadnick, Rong Zhao

- Motivation and Goal:
    - To do unsupervised sentence compression
        - Difficulty in collecting good labeled data
        - Problem in model generalization (i.e. data domain, length)
    - Leverage large machine translation language pair datasets
    - Analyze and study sentence compression results achieved using different language pairs
- Model:
    - Machine Translation Systems
    - Length Control



- Follow-up work based on Jonathan Mallinson, Rico Sennrich and Mirella Lapata "Sentence Compression for Arbitrary Languages via Multilingual Pivoting"

- Experiments
    - Efforts in building a model to target OOV problem caused by domain dissimilarity
        - **fastText word embedding**, wordpiece tokenizer
    - Train multiple NMT systems using eight language pairs
        - length-based hidden cell initialization, **length embedding**
    - Evaluate and analyze the effect of different intermediate languages
        - Human evaluation, ROUGE metric, English fluency test

COMPARISON OF PIVOT LANGUAGES FOR AUTOMATIC SENTENCE COMPRESSION

- Evaluation

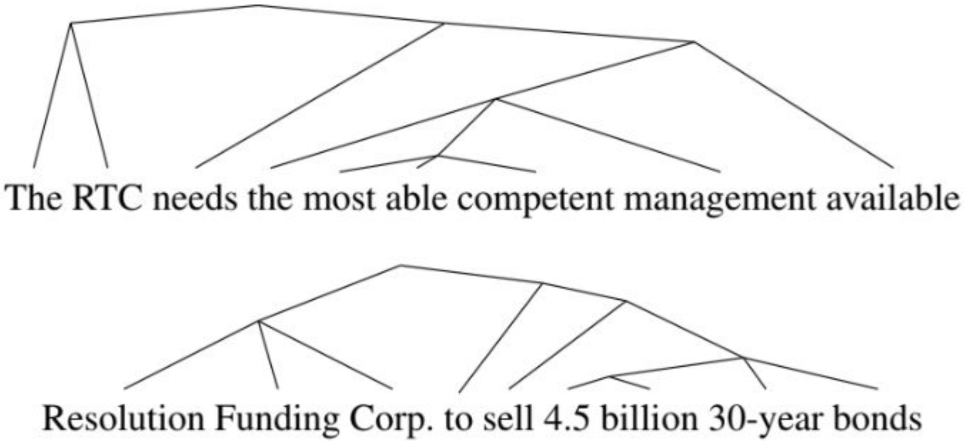
	Dutch	Greek	Italian	Russian	Spanish
Human (Scale 1 to 5)	2.7	2.2	2.9	2.508	3.0
Gigaword ROUGE1 F1	8.2	7.4	9.8	4.3	8.3
MOSS ROUGE1 F1	27.6	29.1	31.2	25.3	33.5
Fluency Test (Scale 1 to 5)	2.85	2.9	2.8	3.48	2.96

- Sample Sentence Compression

Original Sentence	Compressed Sentence
The reason is simple.	simple reason.
Even in just past few years, we've greatly expanded.	in recent years, we've expanded.
Ladies and gentlemen , dear colleagues , it is a great pleasure to welcome here this afternoon the Prime Minister of the United Kingdom , Gordon Brown .	ladies and gentlemen , it is a great pleasure to welcome the prime minister of the united kingdom .
We expect that the Council 's annual report will provide opportunities to establish a dialogue with Parliament aimed at developing a more strategic approach to the **common foreign and security policy** .	we expect the council 's annual report to develop a dialogue with parliament on a more strategic approach to the **cfsp** .
for the latest on mexico , hot off the fax , consult <unk> , a new ##-hour service with tourist information for south of the border .	the UNK UNK UNK UNK .
The organisers of australian fashion week say they will follow the lead of some european countries and keep <unk> models off the catwalks .	next week 's week weekend week

Integrating Pre-Trained Representations into Unsupervised Parsing

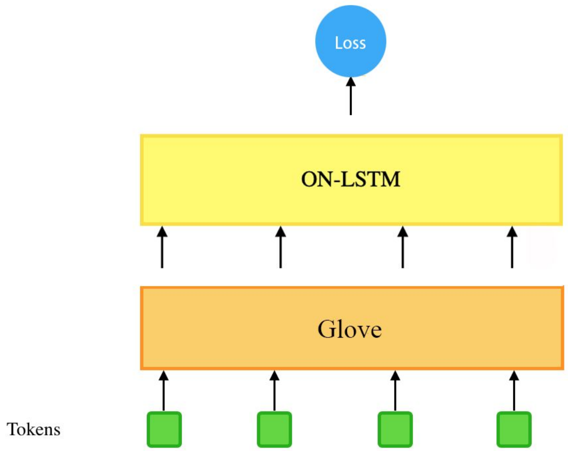
Zhengyang Bian, Yunan Hu, Xinyue Zhang, Bin Zou



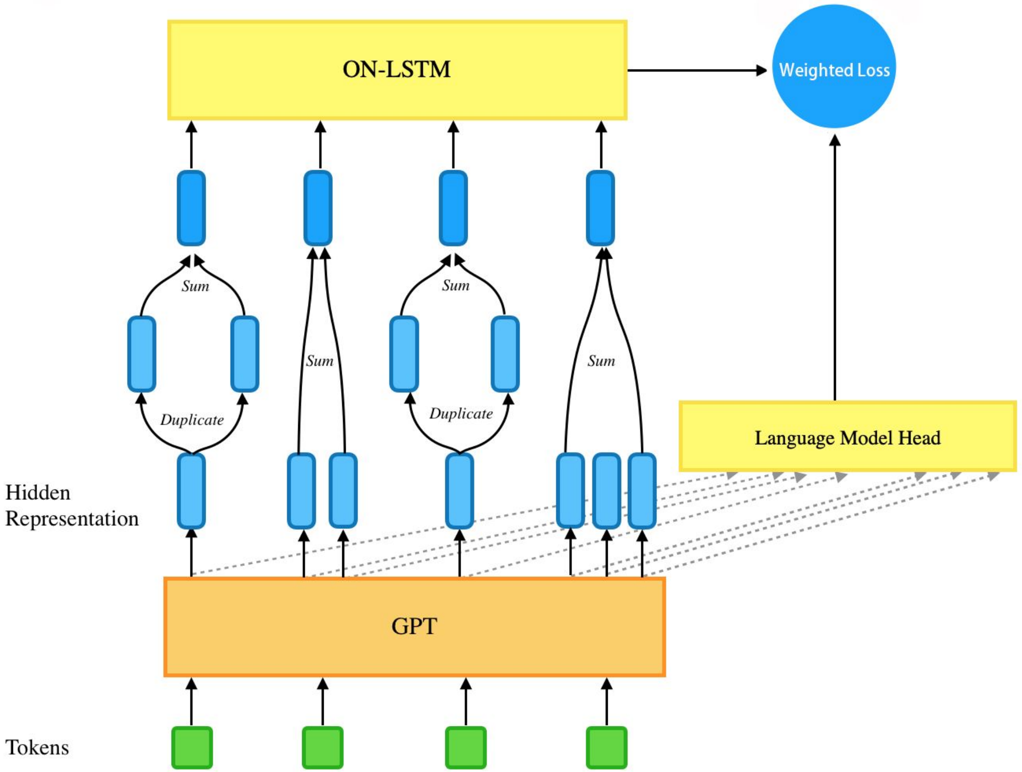
Shen et al., 2018

Model Architecture

ON-LSTM with Glove

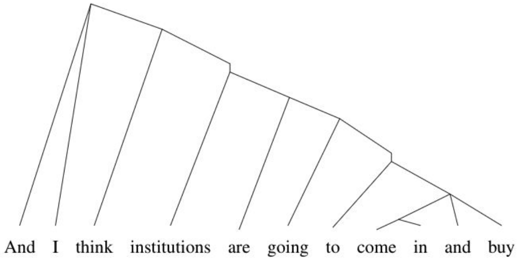


GPT-ON-LSTM

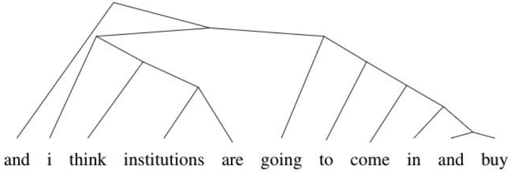


Results

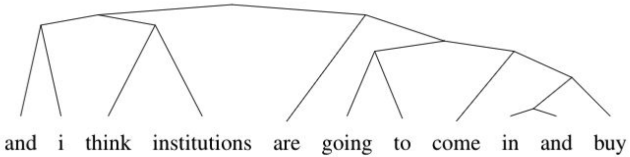
Model	Validation Perplexity	Parsing F1		Depth WSJ	Accuracy on WSJ by Tag			
		WSJ10	WSJ		ADJP	NP	PP	INTJ
ON-LSTM (reported)	-	65.1	47.7	5.8	46.2	61.4	55.4	0
ON-LSTM (reproduced)	58.4	68.9	47.4	5.7	53.0	59.7	55.6	0
ON-LSTM with GloVe Embedding	$\boxed{52.1}$	65.2	$\boxed{48.7}$	5.6	46.6	$\boxed{61.3}$	55.4	0
GPT-ON-LSTM	$\boxed{55.6}$	55.0	41.6	5.7	43.8	54.6	50.2	0



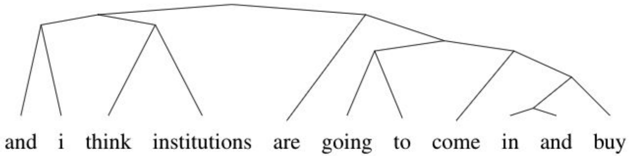
Human



GPT



Glove



Original

String-To-Tree Neural Paraphrase Generation

- Baseline: the Transformer
- Hypothesis: adding in syntactic constraints in the training data can improve baseline's performance on the Paraphrase Generation Task
- Enforces the model to generate sentences that follow a set of grammar rules.

- Transform target sentences into normal linearized tree using PTB parser.

```
ROOT (S (NP (NN john)) (VP (VBZ takes) (NP (DT a) (NN vacation))) (. .))
```

- PTB linearized tree without word level POS tag.

```
ROOT ( S ( NP john ) ( VP takes ( NP a vacation ) ) ( . . ) )
```

Evaluation

	Human	BLEU	METEOR
MSCOCO	36/100	19.8	40.63
MSCOCO + Tree	20/100	5.77	25.90
Twitter	27/100	21.89	42.70
Twitter + Tree	16/100	4.94	25.59
Wikianswers	N/A	23.22	42.94
Wikianswers + Tree	N/A	5.47	25.35

Analysis

- Adding the tree structure led to the generation of more grammarly correct sentences.
- The model focuses more on learning the syntactic structure information rather than semantic understanding and paraphrasing.
- Limited computing resources and time for hyper-parameter tuning due to extremely long sequence length for input.

Text Summarization with Bert and Reinforcement Learning

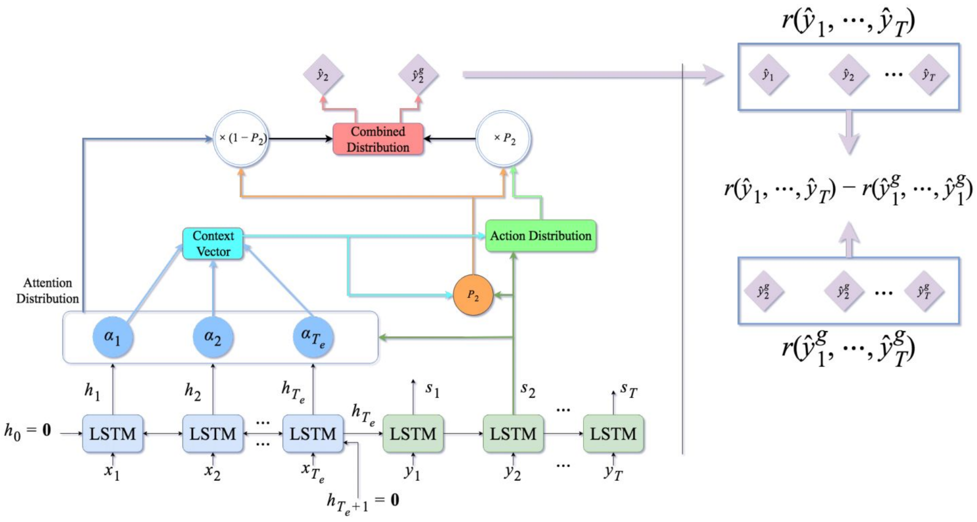
--Krystal Wang, Tia Bi, Sylvie Shao

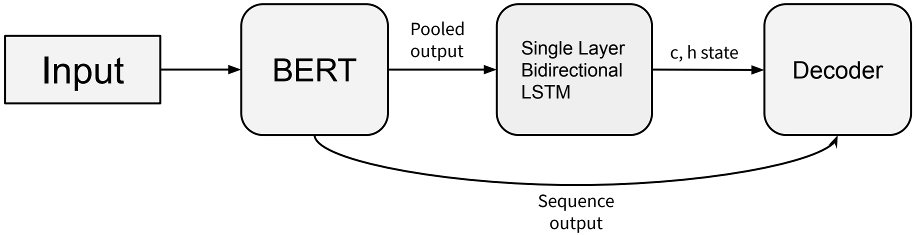
https://github.com/tianyibi/text_summarization

NYU

Project Goal

- Substituting encoder with BERT in Pointer Generator network
- Expect to see increase in performance due to BERT's success in NLU tasks
- Compare results of RL with Pointer Generator network, and BERT with Pointer Generator network



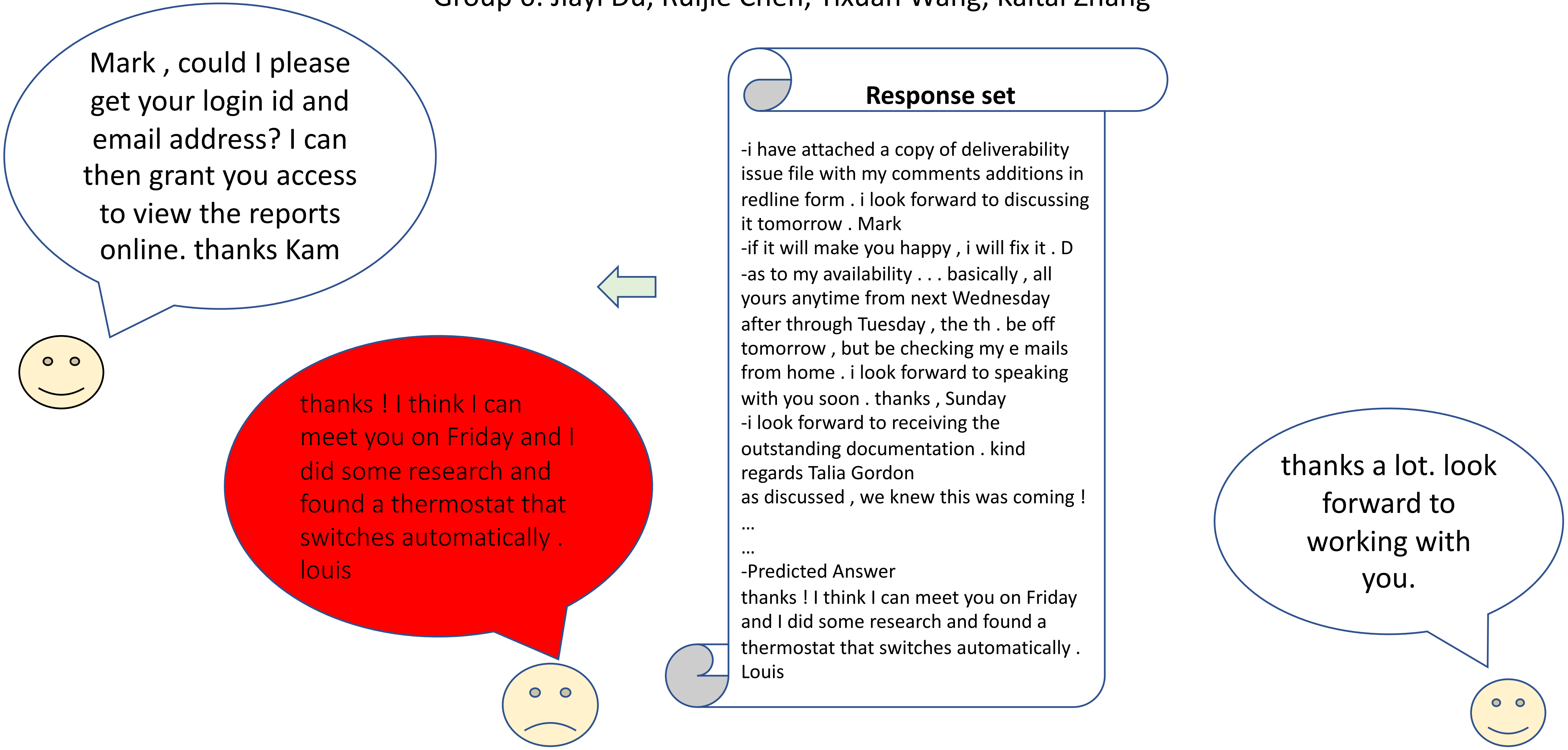


Results

	Before RL		After RL	
Model	**Training Loss**	**ROUGE-1**	**Training Loss**	**ROUGE-1**
RL-Seq2Seq	2.49	32.95	2.50	33.25
BERT (Finetune)	2.12	35.34	2.07	35.51
BERT (Different embedding, Finetune)	2.25	33.84	2.03	34.01

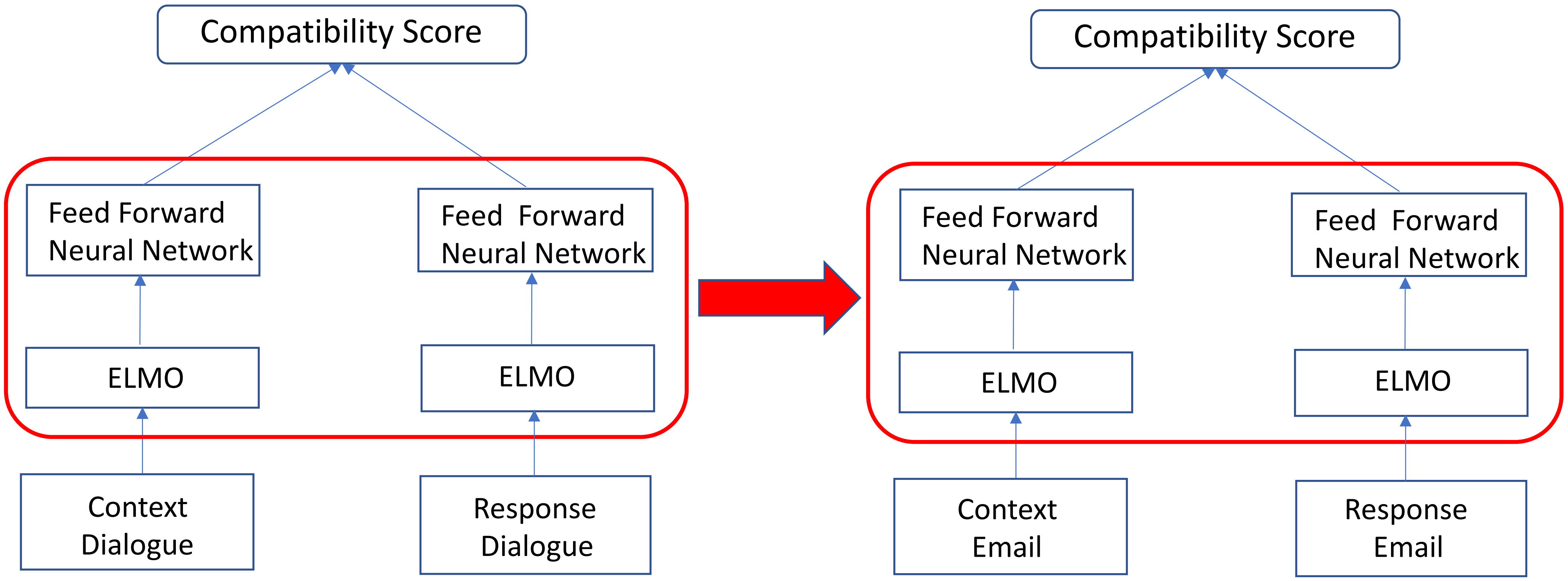
Use of Transfer Learning to Improve Automatic Email Reply Quality

Group 6: Jiayi Du, Ruijie Chen, Yixuan Wang, Kaitai Zhang



Use of Transfer Learning to Improve Automatic Email Reply Quality

- Model

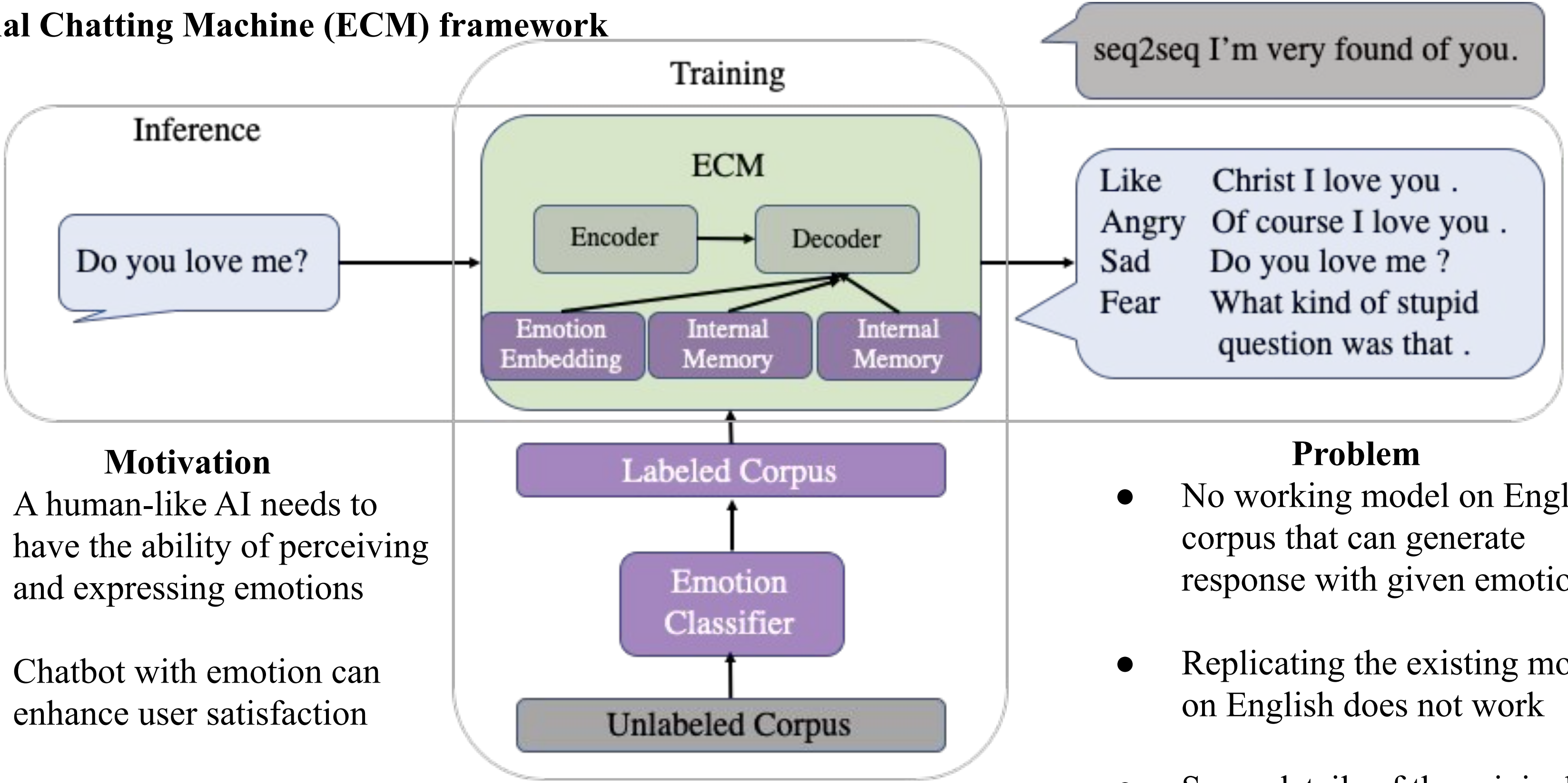


First Train on the Dialogue Dataset

Then Fine-tune on the Email Dataset

Chatbot with Emotional Chatting Machine

Emotional Chatting Machine (ECM) framework



Motivation

- A human-like AI needs to have the ability of perceiving and expressing emotions
- Chatbot with emotion can enhance user satisfaction

Problem

- No working model on English corpus that can generate response with given emotion
- Replicating the existing model on English does not work
- Some details of the original ECM framework are flawed

Bofei Zhang, Yuxuan He, Shaoling Chen, Yu Cao

Hypothesis and Novel Methods

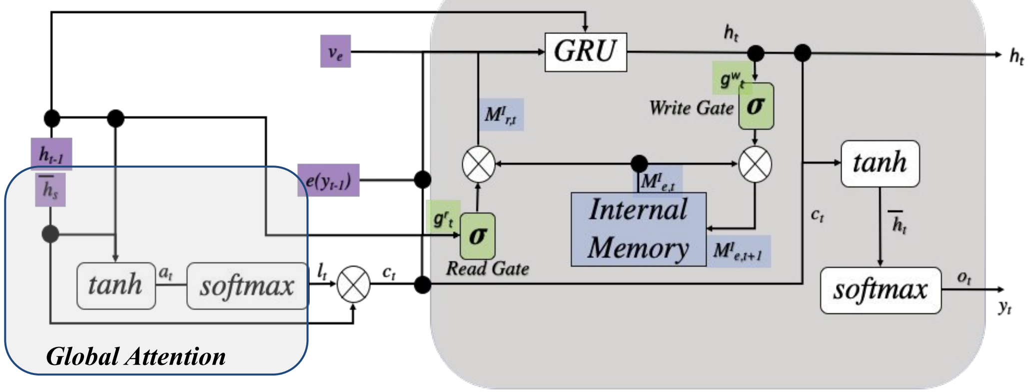
Hypothesis

- The idea of ECM can be used in English dataset
- ECM can generate more emotional and grammatically coherent response than seq2seq model

Novel methods

- Adopt **global attention** mechanism (Luong et al., 2015)
- Simplify **external memory** module
- Add static emotion embedding to **internal memory** module

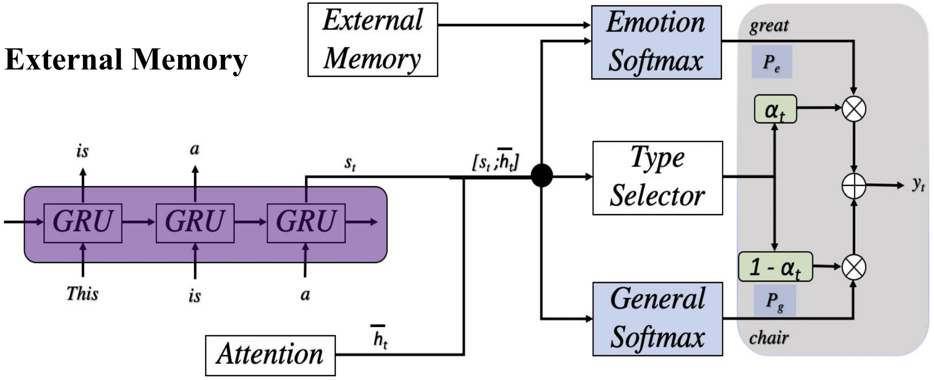
Internal Memory



$$L(\theta) = -\sum_{t=1}^{m} \boldsymbol{p}_t \log(\boldsymbol{o}_t) - \sum_{t=1}^{m} q_t \log(\alpha_t) + \| \boldsymbol{M}^I_{e,m} \|,$$

(Zhou et al.,2018)

External Memory



Bofei Zhang, Yuxuan He, Shaoling Chen, Yu Cao

NYU

Results and Conclusion

Multiturn Conversation	Seq2seq	Joy	Angry
Human	where are you from ?		
Bot	i don't know ...	I'm from shanghai . And you?	I'm from california .
Human	are you from shanghai ?	same place ! do you like shanghai ?	is california good ?
Bot	yes .	That's **fantastic** . Thank you .	oh **god** . That's not.
Human		Let's go to a bar in shanghai !	what happened to you in california ?
Bot		have a lot of drink !	**never** face a business

Post	Model		Response
Do you love me?	Seq2Seq		I'm very fond of you .
	ECM	Joy	christ I **love** you .
		Anger	of course I **love** you .
		Sadness	do you **love** me ?
		Fear	what kind of **stupid** question was that .

	Machine Evaluation	Human Evaluation			
Model	**Perplexity**	**Content Relatedness**	**Grammar**	**Emotion Intensity**	**Emotion Accuracy**
Seq2Seq	1.28	3	3.88	2.67	37%
ECM w/o Ememory	1.37	2.48	3.57	2.42	40%
ECM	1.32	2.81	**4.11**	**2.74**	**40.70%**
Training Data	-	3.09	4.39	2.84	51.60%

Conclusion

- The concept ECM can be applied on English corpus.
- ECM outperforms seq2seq model in emotion accuracy and grammatical coherence.
- Emotions are hard to model

Bofei Zhang, Yuxuan He, Shaoling Chen, Yu Cao

Checkout the website to evaluate our model! https://bofei.shinyapps.io/ECM-Experiment/

ORDERING ISSUES FOR OUTPUT SETS IN UNIFIED SEQUENCE-TO-SEQUENCE OPEN INFORMATION EXTRACTION MODELS

TINGYAN XIANG (TX443)
TIANYU WANG (TW1682)

OPEN INFORMATION EXTRACTION

	Chinese	English
Sentence	唐娜·凯伦（Donna Karan）出生于纽约长岛，对纽约这个世界大都会有着一份特殊的感悟。	Donna Karan (唐纳·凯伦) was born in Long Island, New York. She has a special comprehension of New York a cosmopolitan city.
Facts	(唐娜·凯伦 $ _ $ Donna Karan) (唐娜凯伦 $ 出生于 $ 长岛) (唐娜·凯伦 $ 对 X 有着 Y $ 纽约 $ 一份特殊的感悟)(长岛 $ IN $ 纽约) (纽约 $ ISA $ 世界大都会)	(Donna Karan, _ , 唐纳·凯伦) (Donna Karan, born in, Long Island) (Donna Karan, has a X of Y, special comprehension, New York) (Long Island, IN, New York) (New York, ISA, cosmopolitan city)

Model: machine-translation-like architecture

- Seq2Seq Model with copy mechanism

Issue & Purpose:

- Does fact ordering impact our model performance?
- In practice, what’s the “best” order for training?

ORDER IMPACT

Test Score	Precision	Recall	F_1-score
baseline	39.09	26.88	29.98
appearance	38.73	36.16	36.64
reverse	37.69	33.61	34.70
last3	38.24	33.23	34.60

Conclusion:

- Output ordering impacts model performance in practice

Ordering Choices:

- Based on some prior knowledge
- Pick the “best” order automatically

HOW TO CHOOSE THE "BEST" ORDER FOR TRAINING

Algorithm 1 Searching Order

t=0, T_1=permutation steps, T=total steps
while $t < T$ **do**
  encoding
  if $t < T_1$ **then**
    choose a permutation order π_c
  else $\{t \geq T_1\}$
    calculate $P(Y_{\pi_l}|X), l = 1, \cdots, n!$ through running decoder
    pick an order π_c according to a distribution proportional to $P(Y_{\pi_l}|X)$
  end if
  decoding based on the chosen order π_c
end while

Test Score	Precision	Recall	F_1-score
baseline	39.09	26.88	29.98
appearance	38.73	36.16	36.64
reverse	37.69	33.61	34.70
last3	38.24	33.23	34.60
permutation	40.67	25.94	29.56
search-20	39.15	29.49	31.85
search-100	40.32	31.79	33.94
search-200	38.46	33.40	34.57

Over 80% samples learn the appearance order as the best

Exploring ways to improve Coreference Resolution

"**Yada, Priyank, and Omkar** like learning about **NLP.** **They** find **it** fascinating"

- Current SOTA (Lee et al. 2018) uses an end-to end neural network model
- Can be broken into Span Identification & Classification
- Brief overview:
  - Use ELMo (frozen) + Bidirectional LSTM (per sentence) to create contextual word embeddings
  - Use attention to score words in span to get top k spans
  - Neural Net based similarity matrix between antecedents and span
  - Treating span-antecedent identification as a classification task

Yada Pruksachatkun, Priyank Pathak, Omkar Damle

Ablation Study of Lee’s model

Model	F1 Results (CoNLL)
Baseline	73.50
w/o genre	71.48
w/o span width embedding	73.15
w/o speaker embedding	72.39
w/o char embedding	72.96

Where the model goes wrong:

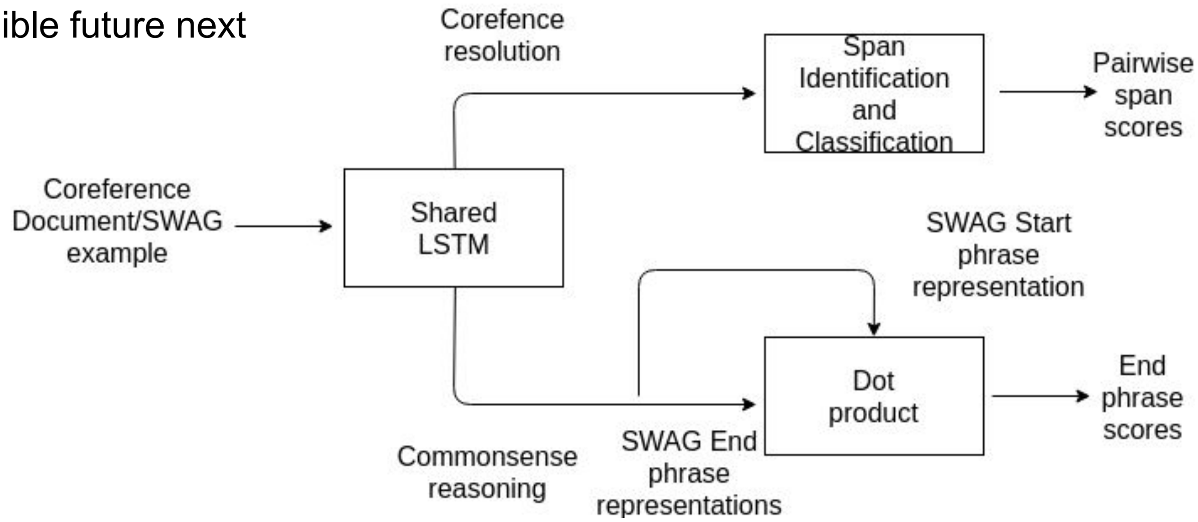
- Doesn’t do well with spans that are further from each other in the source text.

- Bias towards pronoun identification

Yada Pruksachatkun, Priyank Pathak, Omkar Damle

Multitask Training on SWAG

SWAG consists of 113K multiple choice questions. Each SWAG example has 4 entries, the correct one and 3 incorrect ones (one context and 4 possible future next sentences).



Lee's model with BERT (F1)	Lee's model with BERT and SWAG multitask training (F1)
82.4	**84.0**

Yada Pruksachatkun, Priyank Pathak, Omkar Damle

Multi-Label Emotion Classification in English Poetry using Song Lyrics and a Dual Attention Transfer Mechanism

Emotion classification captures nuance present in text that is unaddressed by sentiment classification

Multi-Label Emotion Classification in **English Poetry using Song Lyrics** and a Dual Attention Transfer Mechanism

GENIUS

My tea's gone cold I'm wondering why / I got out of bed at all The morning rain clouds up my window / And I can't see at all / And even if I could it'd all be gray / But your picture on my wall / It reminds me that it's not so bad / …

☐ Positive ☑ Negative

amazon mechanical turk

r/OCPoetry

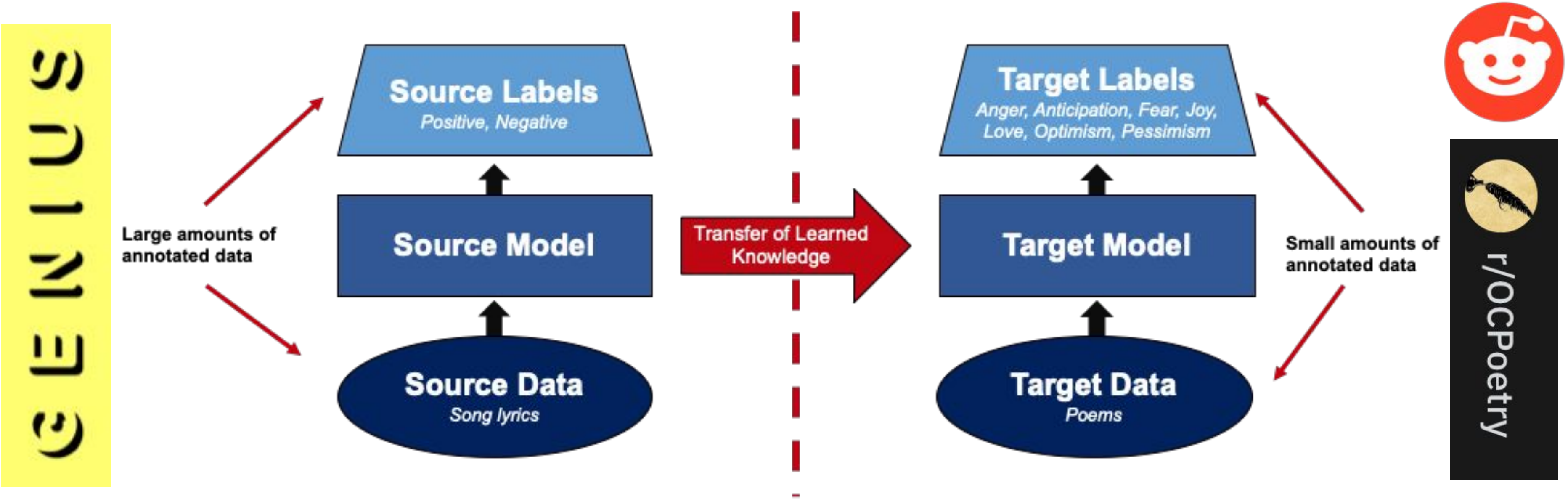
It's about time she said / But that's not how she meant it / Its about the timing / When the galaxies outside our solar system align / To form a perfect map of where we've been / Or more perfectly where we could be / But I can only see so far / And you can only drift so close / So you orbit me like the stars / Always out of reach

Which TWO emotions does this poem invoke? (Please only select two.)

☐ Anticipation ☐ Anger ☐ Fear ☐ Optimism ☐ Joy ☑ Love ☑ Pessimism ☐ Sadness

Transfer learning repurposes a model trained on a separate task to enhance performance

Multi-Label Emotion Classification in English Poetry using Song Lyrics and a **Dual Attention Transfer Mechanism**



Baseline Accuracy: 40.4%

DATN Accuracy: 30.3%

Visual Question Answering with Transfer Learning for Question Encoding

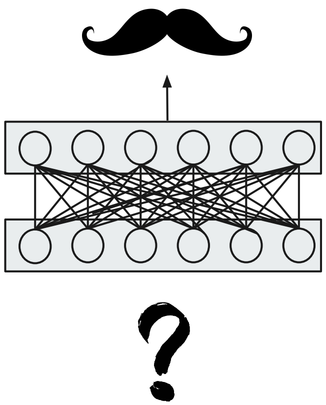
Stephen Carrow Chris Rogers Isaac Haberman Hollis Nymark

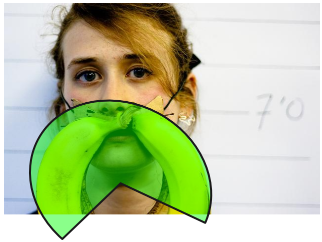
Q: What is the mustache made of?
A: Bananas

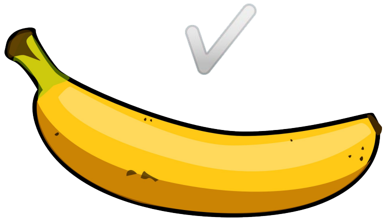
Visual Question Answering with Transfer Learning for Question Encoding

Stephen Carrow Chris Rogers Isaac Haberman Hollis Nymark

Better Question Representation ⇒ Better Visual Information ⇒ Better Answers







Visual Question Answering with Transfer Learning for Question Encoding

Stephen Carrow Chris Rogers Isaac Haberman Hollis Nymark

Method	Overall Accuracy
MCB - Baseline	61.96
MCB	59.52
MCB + GLoVE	60.56
MC-ELMo	59.89
MC-BERT	59.45

Comparison of VQA architectures using different question encoders

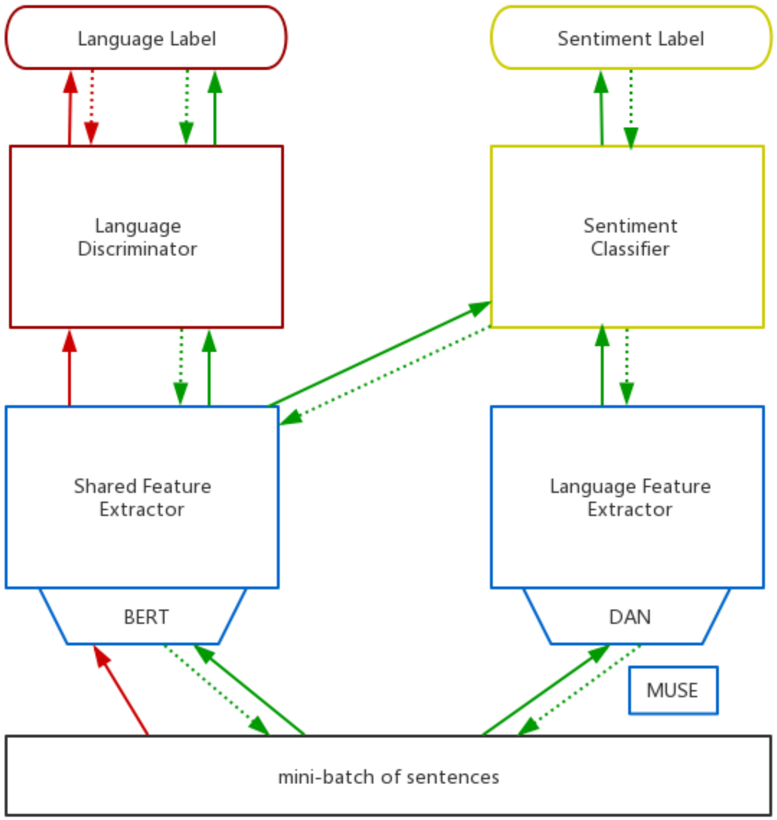
- MCB - Baseline is the published result.
- Our results lagged, but allow internal comparisons.
- MCB+GloVE shows some improvement.
- The more complex models didn't do as well.

Cross-Lingual Sentiment Classification Using Multinomial Adversarial Networks

- Problem CLSC with varying amount of data from several languages
- Goal Improve the overall classification performance across all languages

Zimo Li, Jiahui Li

Model Architecture



Dataset

Amazon Customer Reviews Dataset

Selected domains: Book, DVD, Music, Mobile

Languages: English, French and German

Results

	Fr	En	De	Avg.
Domain-Specific Models Only				
book	81.84	84.08	79.08	81.67
dvd	87.88	88.69	89.67	88.74
mobile	87.05	84.89	86.70	86.21
music	85.8	85.95	85.85	85.87
Shared Models Only				
book	81.58	84.61	79.61	81.93
dvd	88.0	88.54	89.36	88.64
mobile	88.18	85.57	86.48	86.74
music	85.75	85.85	86.4	86
Shared Models with Discriminator				
book	**82.5**	83.95	79.74	**82.06**
dvd	87.58	88.60	89.15	88.44
mobile	88.30	85.11	86.82	86.74
music	85.95	86.4	86.1	86.15
Shared-Private Models				
book	**83.82**	84.08	**80.79**	**82.90**
dvd	88.02	88.48	89.15	88.54
mobile	87.84	85.11	86.25	86.40
music	86.2	85.5	87.35	86.35

Table 1: Domain-Specific Results

	Fr	En	De	Avg.
Domain-Specific Models Only				
MAN-Baseline	86.8	87.38	87.14	87.1
MAN-Bert	87.09	85.86	86.75	86.57
Shared Models Only				
MAN-Baseline	87.08	87.33	86.86	87.09
MAN-Bert	**88.21**	**90.14**	86.75	**88.36**
Shared Models with Discriminator				
MAN-Baseline	87.29	87.25	86.86	87.13
MAN-Bert	**89.13**	**89.94**	**88.44**	**89.17**
Shared-Private Models				
MAN-Baseline	87.03	87.44	86.88	87.12
MAN-Bert	86.83	87.1	87.01	86.98

Table 2: Domain-Invariant Results

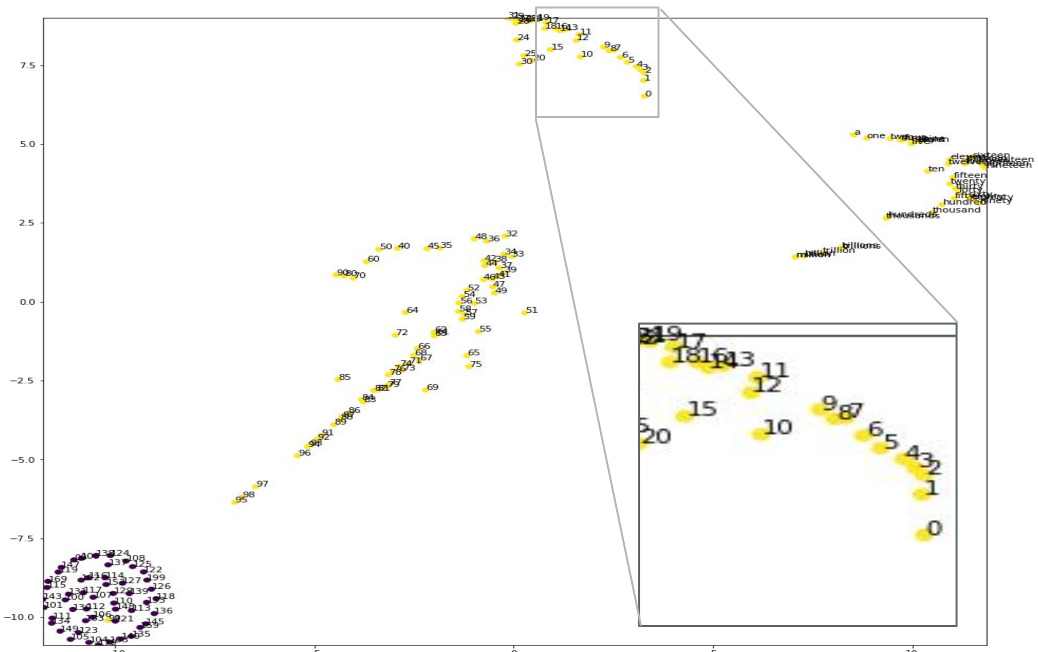
JP Park, Grace Han, Yanchao Ni, Mingsi Long

Breaking Numerical Reasoning in NLI

- SoTA NLI models fail on our adversarial test set, especially neutral category

		SNLI original	adversarial without addition				adversarial with addition			
Model	Embedding	all	all	entail	cont	neutral	all	entail	cont	neutral
ESIM	GloVe	87.46	24.88	91.32	88.44	8.72	54.31	98.80	97.10	0
BERT	N/A	90.44	22.48	89.04	89.41	5.69	41.72	66.47	97.10	0

- GloVe Embedding has unclear relationship between number words
- Word analogies fail as well (e.g. **Three** to **two** is **eight** to ~~four~~)
  seven



Hypothesis + Methods

"Augmenting data and/or new embedding will resolve the problem."

1. Entailment

   a. ∄ **Addition**: For pairs where $num_p = num_h$, iterate from 2~10 and replace num_p and num_h while maintaining $\boldsymbol{num_p = num_h}$

   (2 boys are running / 2 kids are moving)
   ⇒ *(3 boys are running / 3 kids are moving)*

   b. ∃ **Addition**: For pairs that have two numerical words in premise and one in hypothesis where $num_{p1} + num_{p2} = num_h$, iterate from 2~10 and replace with new values while maintaining $\boldsymbol{num_{p1} + num_{p2} = num_h}$

   (There are 2 dogs and 3 cats / 5 animals)
   ⇒ *(There are 3 dogs and 4 cats / 7 animals)*

2. Neutral (Same as entailment except P < H)

   a. *(2 boys are running / 2 kids are moving)*
   ⇒ *(2 boys are running / 3 kids are moving)*

   b. *(There are 2 dogs and 3 cats / 5 animals)*
   ⇒ *(There are 3 dogs and 4 cats / 8 animals)*

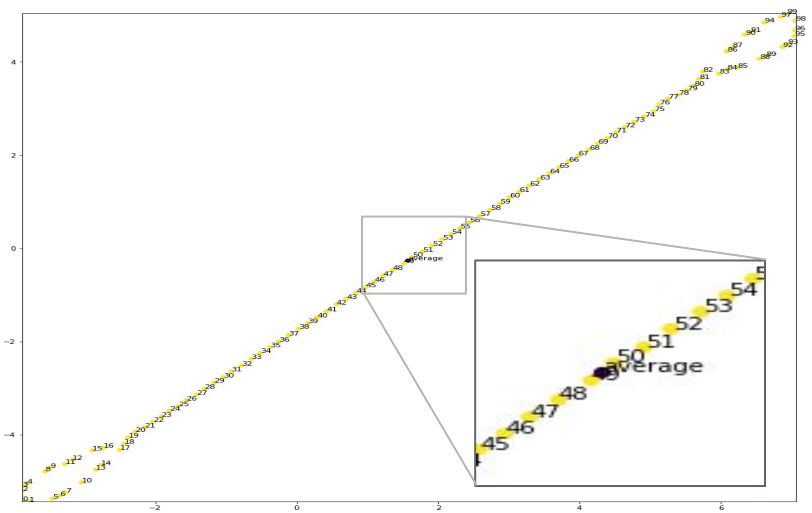
3. Contradiction

   a. Premise ↔ Hypothesis

   (Two boys are singing / Two boys sleeping)
   ⇒ (Two boys sleeping / Two boys are singing)

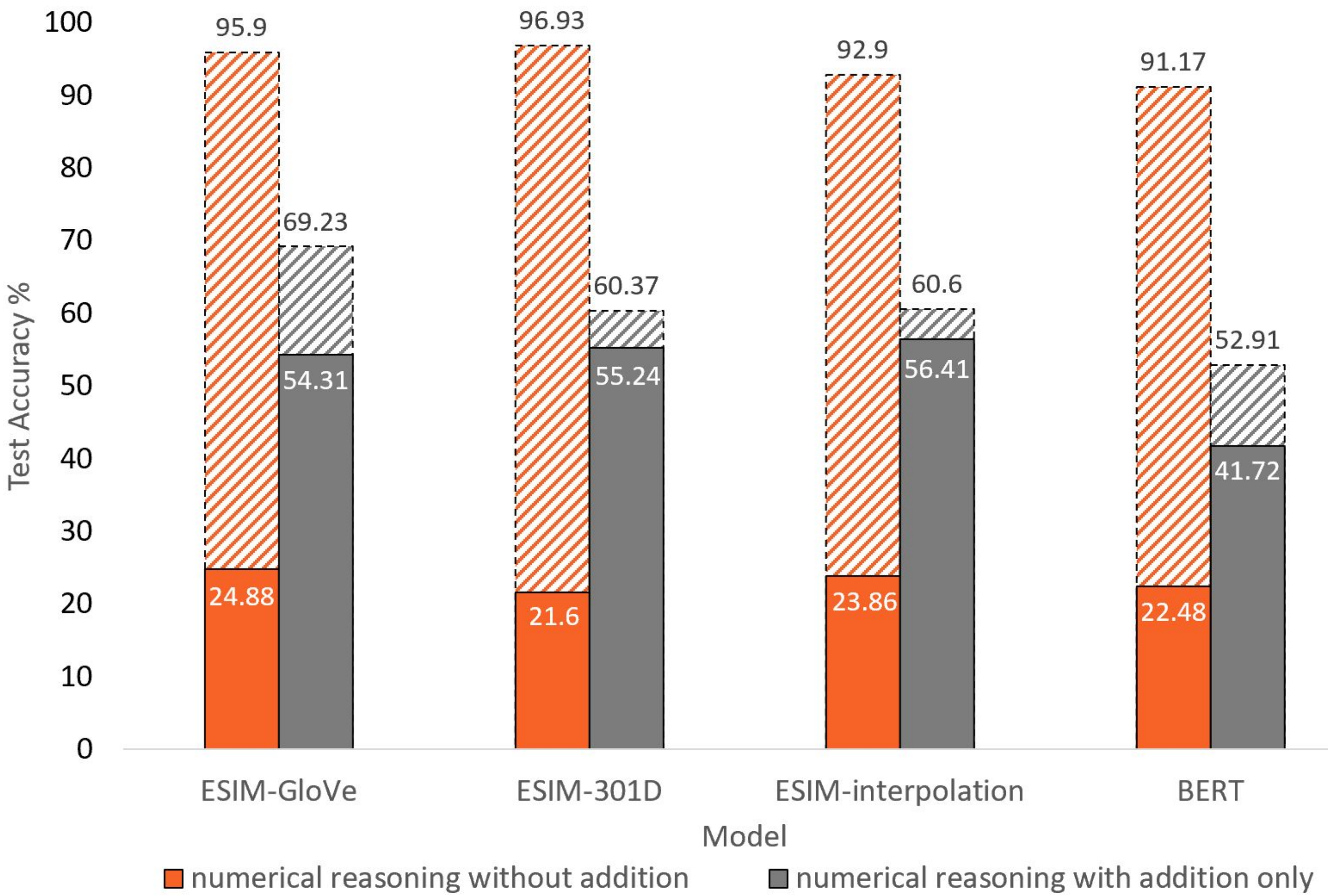
   b. Replace object to an antonym

   (Two boys are singing / Two kids singing)
   ⇒ (Two boys are singing / Two adults singing)



Modified Embedding has linear relationship and all analogies succeed.

Results + Analysis



We speculate that the current architectures cannot do complicated numerical reasoning beyond simple pattern matching since we excluded data and embedding.

Automated Lyric Annotation

Jay is announcing his return to the rap scene after being absent

Allow me to re-introduce myself

My name is HOV, OH, H-to-the-O-V

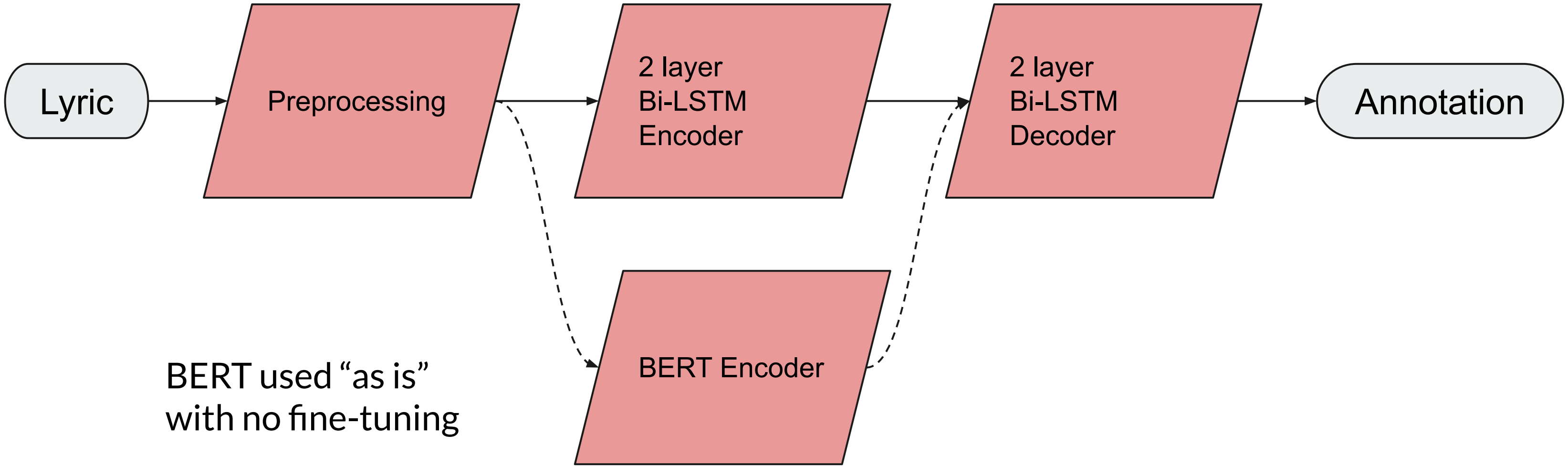
I used to move snowflakes by the OZ[1]

Exodus 6:3: “And I appeared unto Abraham, unto Isaac, and unto Jacob, by the name of God Almighty, but by my name JEHOVAH was I not known to them.”

To “move snowflakes by the OZ” is to deal cocaine (successfully)

1. Carter, Shawn. Public Service Announcement (2003). *The Black Album*

Our Model: BERT Encoder



BERT used “as is”
with no fine-tuning

Results and Annotations

Lyric	wine drunkenness over the rooftops	Fluency	Information
Human	hipster parties on the rooftops listening to bird and sipping some red	5.00	1.29
Retrieval	a lot of ARTIST's music is about small meaningful memories which have played a large part (emotionally) in trevor power's life, simple things like watching fireworks explode from a rooftop	4.71	3.86
Seq2Seq	wine is often used as a metaphor for alcohol	4.00	3.86
BERT	this is a play on words. he's saying that he's got a lot of money	4.57	1.86
Lyric	lost your balance on a tightrope. lost your mind trying to get it back		
Human	many people found this act to be kanye's last straw. he was heralded as crazy, people theorized west had lost complete control.	4.43	4.00
Retrieval	she is saying that even if you fall, you can still get back up and continue on. kanye is no different	4.43	3.33
Seq2Seq	a reference to the song "life's a b****"	5.00	1.00
BERT	this is a play on words. he's saying that he's got a lot of money	4.43	3.43

Overcoming Catastrophic Forgetting in Pre-training Language Models

{Ethan Perez, Ananya Harsh Jha}

Elastic Weight Consolidation: Training Objective

Fine-tuning Loss and Params

Regularization Strength

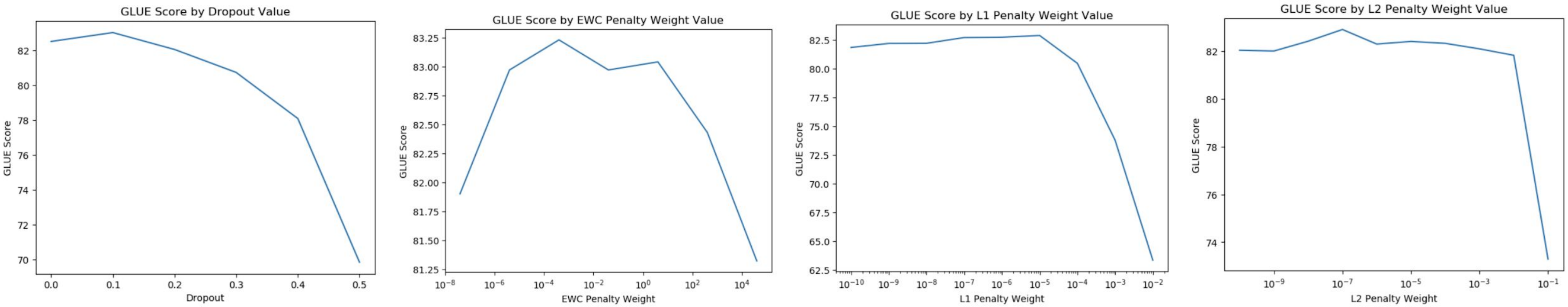
Pre-training Loss and Params

$$\mathcal{L}_F(\theta_F) + \sum_i \frac{\lambda}{2} \underbrace{\left(\frac{d\mathcal{L}_P(\theta_P)}{d\theta_{P,i}}\right)^2}_{\text{Parameter Importance}} \underbrace{(\theta_{F,i} - \theta^*_{P,i})^2}_{\text{Parameter Distance}}$$

Results

Coefficient Selection	*Method*	*MNLI-(m/mm) 392k*	*QQP 363k*	*QNLI 108k*	*SST-2 67k*	*CoLA 8.5k*	*STS-B 5.7k*	*MRPC 3.5k*	*RTE 2.5k*	*Average*
-	**BERT$_{\mathrm{BASE}}$**	83.72/84.17	89.52	88.43	93.00	58.06	89.94	89.95	71.48	83.04
Best Overall	$\updownarrow$ **Dropout**	83.72/84.17	89.52	88.43	**93.00**	58.06	**89.94**	89.95	71.48	83.04
	$\uparrow$ **L1**	**84.44/84.52**	87.61	88.73	92.43	**61.03**	89.11	89.08	70.40	82.86
	$\uparrow$ **L2**	84.11/84.41	90.08	**88.76**	**93.00**	60.52	89.30	89.76	67.51	82.90
	$\uparrow$ **EWC**	83.71/83.94	**90.11**	88.68	92.43	58.54	89.12	**90.22**	**72.92**	**83.23**
Task-Specific	$\updownarrow$ **Dropout**	83.84/84.17	89.70	88.66	**93.46**	59.72	**90.05**	89.95	71.48	83.38
	$\uparrow$ **L1**	**84.44/84.52**	90.06	**88.96**	93.35	**61.10**	89.35	89.16	71.12	83.45
	$\uparrow$ **L2**	84.32/84.47	90.09	89.13	93.23	60.52	89.33	89.83	71.84	**83.55**
	$\uparrow$ **EWC**	83.97/84.17	**90.20**	88.92	92.66	59.82	89.14	**90.22**	**72.92**	83.49

Performance by Regularization Strength



(a) Dropout (b) EWC (c) L1 (d) L2

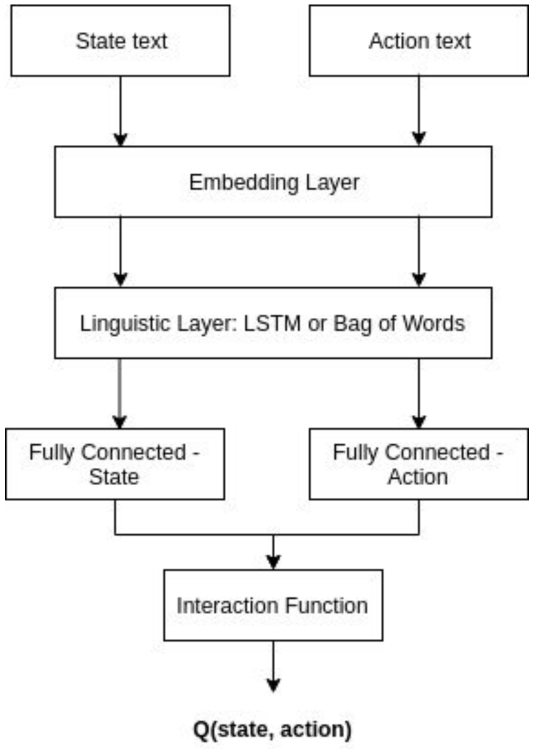
On each plot: Right is More Regularization

Transfer of Reinforcement Learning in a Natural Language Action Space

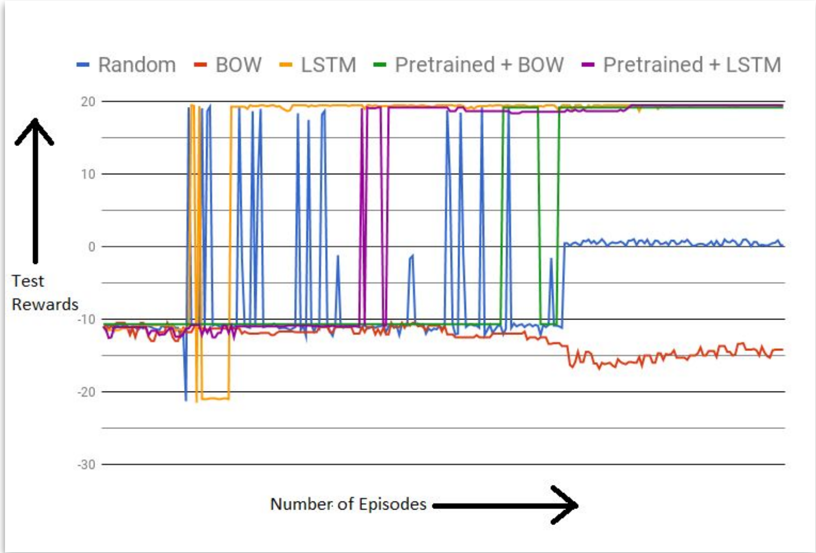
- Natural Language Action Space? - Text-based Games - PyFiction (Zelinka et. al.)

- Why is transfer interesting?
    - “What can you do with a sentence if you know its meaning?”
    - Deeper understanding, better generalisation

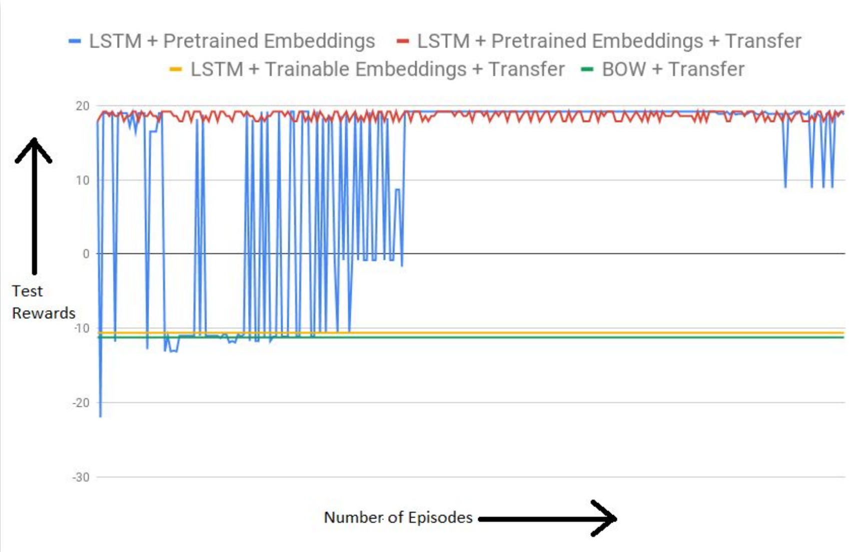
But how?


State text
Action text
Embedding Layer
Linguistic Layer: LSTM or Bag of Words
Fully Connected - State
Fully Connected - Action
Interaction Function
Q(state, action)

What do we observe?



Single game setting



Transfer setting

Key-Value Memory Network Model

Attentive Query using Knowledge Base

Wenting Qi
Zhiyuan Wang
Yiyi Zhang
Weicheng Zhu

Question Classification

Background

Task: Medical Question Classification

Data: **1.6M** Medical QA logs crawled from HealthTap

Input: Questions including patients' descriptions of symptoms

Output: **225** question categories

Data Exploratory:

Top **20%** minority labels consist of only **0.4%** of the overall data population

Example of Records with Minority Labels

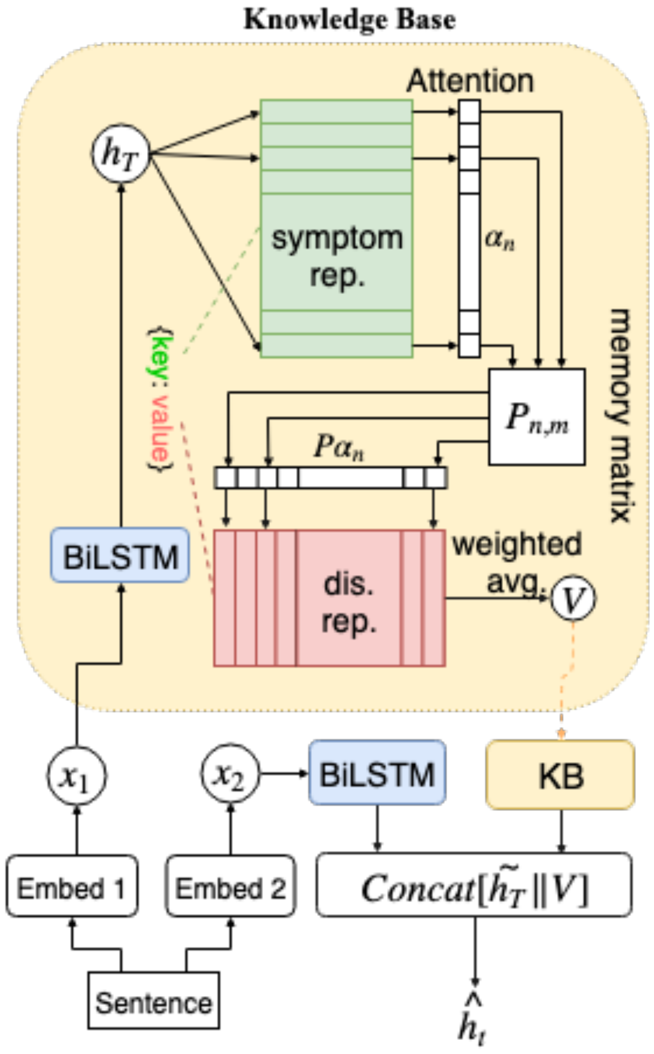
Q: Had gastritis what to do?

Category: stomach

Issues with Such Data:

Data sparsity for rare but sometimes valuable labels

Attentive **K**nowledge-**B**ased **M**emory **N**etwork (AKB-MN)



Model Structure

- Bi-LSTM
- Knowledge Base
  - Bi-LSTM
  - Attention Mechanism
  - Memory Matrix

Evaluation

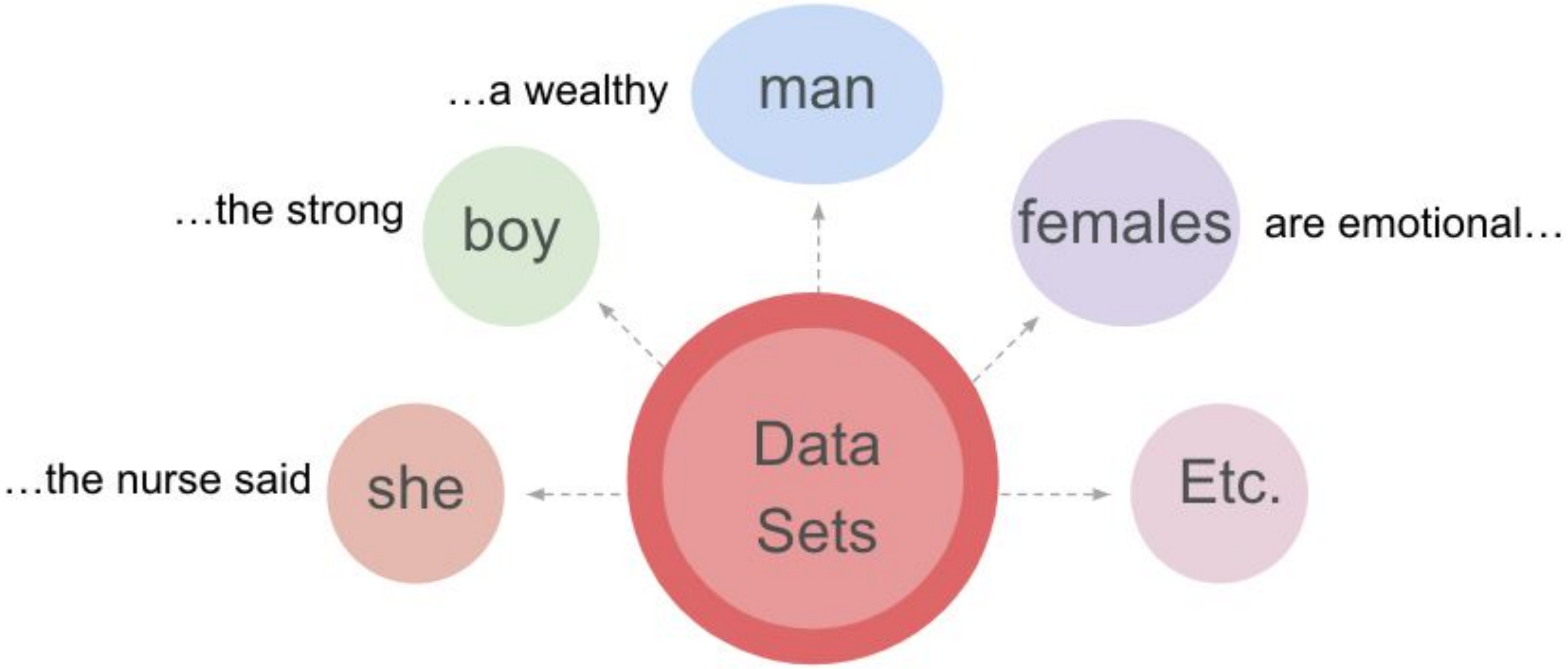
Performance on Minority Labels

- AKB-MN improves overall accuracy by 0.9%
- Improves over 50% of the minority labels' accuracy

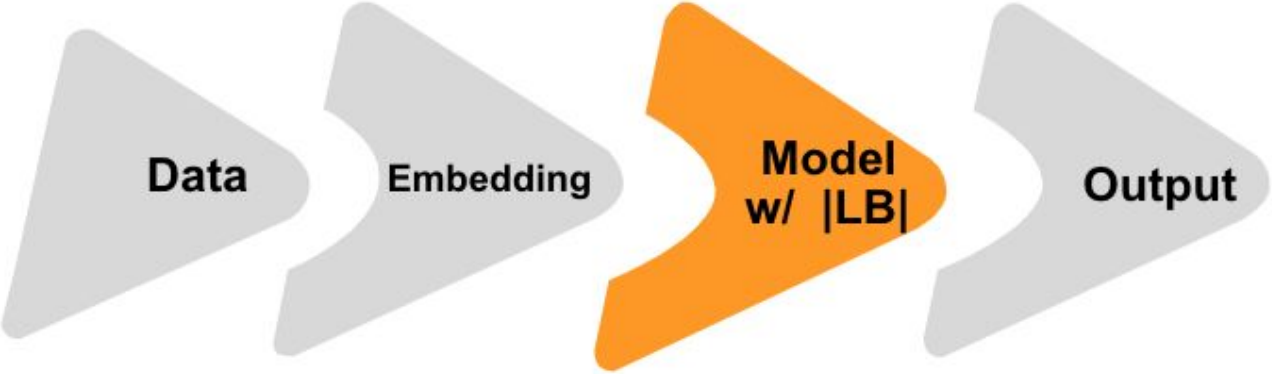
	Label	Count	KB Accuracy	Baseline Accuracy
1	tone	15	0.466667	0.333333
5	nausea	25	0.560000	0.400000
6	charley horse	40	0.775000	0.675000
10	amalgam filling	93	0.698925	0.537634
12	stomach	100	0.640000	0.390000
14	body	144	0.576389	0.347222
18	carbidopa levodopa	188	0.585106	0.457447

Test Performance Comparison between Baseline Model and AKB-MN

Sources of Gender Bias in Natural Language Datasets



Methodology



Cross-entropy

$$L^{CE}(t) = -\sum_{w\in V} y_{w,t}\log\left(\hat{y}_{w,t}\right)$$

Our add-on

$$L^{B}(t) = \frac{1}{G}\sum_{i}^{G}\left|\log\frac{\hat{y}_{f_i,t}}{\hat{y}_{m_i,t}}\right|$$

$$L = \frac{1}{T}\sum_{t=1}^{T} L^{CE}(t) + \lambda L^{B}(t)$$

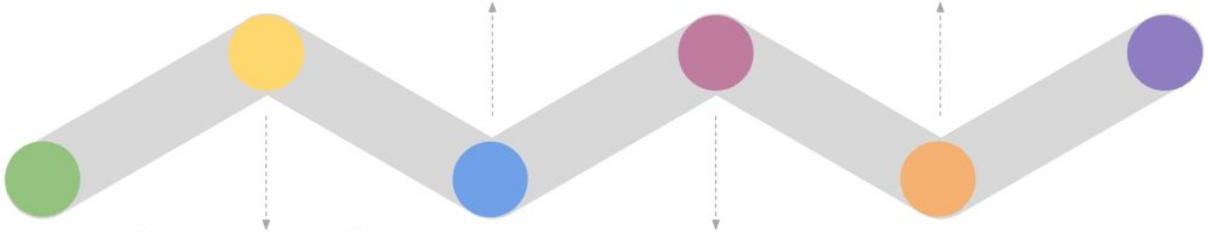
Bias Evaluation

$\overbrace{\text{The } \textbf{doctor} \text{ is a }|}^{s}$ $\overbrace{\textbf{man}}^{t_1}$ / $\overbrace{\textbf{woman}}^{t_2}$ $\log\frac{P(t_1|s)}{P(t_2|s)}$

Occupation Bias Conditioned on Occupation

$$EB_d = \sum_{o\in O}\sum_{i}^{G}\left|\|E(o)-E(m_i)\|_2 - \|E(o)-E(f_i)\|_2\right|,$$

Word Embedding Bias



Co-occurrence Bias

$$B^N = \frac{1}{N}\sum_{w\in N}\left|\log\frac{c(w,m)}{c(w,f)}\right|$$

$$B^N_c = \frac{1}{N}\sum_{w\in N}\left|\log\frac{P(w|m)}{P(w|f)}\right|$$

Occupation Bias Conditioned on Gender

$\overbrace{\textbf{He} \text{ is a }|}^{s_1}$ / $\overbrace{\textbf{She} \text{ is a }|}^{s_2}$ $\overbrace{\textbf{doctor}}^{t}$ $\log\frac{P(t|s_1)}{P(t|s_2)}$

Results

Model	B^N	B^N_c	GR	$Ppl.$	$CB\|o$	$CB\|g$	EB_d
Baseline	0.531	0.282	1.415	117.845	1.447	97.762	0.528
REG	0.381	0.329	1.028	**114.438**	1.861	108.740	0.373
CDA	0.208	0.149	1.037	117.976	0.703	56.82	0.268
$\lambda_{0.5}$	0.312	0.173	1.252	120.344	**0.000**	1.159	0.006
λ_1	0.218	0.153	1.049	120.973	**0.000**	0.999	0.002
λ_2	0.221	0.157	1.020	123.248	**0.000**	0.471	**0.000**
$\lambda_{0.5}$ + CDA	**0.205**	**0.145**	**1.012**	117.971	**0.000**	**0.153**	**0.000**

Our model reduced

Conclusion

- Debiasing the model with bias penalties in the loss function is an effective method.
- This method is powerful and generalizable to downstream NLP applications.
- Geometric debiasing of the word embedding is not a complete solution for debiasing the downstream applications.

VQA via Reason Prediction

Mihir Rana and Kenil Tanna

Question: Are [person1] and [person2] happy to be here?

Rationale: I think that...

a) [person1] looks distressed, not at all happy. **14.6%**
b) [person2] is in an argument with [person1] which does not look like it is resolving. **0.7%**
c) Both their expressions are unhappy and unimpressed with their surroundings, and they look out of place. **80.8%**
d) They both are looking down and emotional. **3.9%**

Answer: So the answer is...

a) Yes, they will spend the night here. **0.1%**
b) No, neither of them is happy, and they want to go home. **77.0%**
c) No [person1] and [person2] are not happy, they seem scared. **20.5%**
d) Yes, [person1] and [person2] are happy. **2.4%**

Example from VCR dataset (Zellers et al., 2019) modified for our approach

Code: https://github.com/ranamihir/visual_commonsense_reasoning

VQA via Reason Prediction

Mihir Rana and Kenil Tanna

Accuracy Comparison VCRSmall-Test

- No one-size-fits-all model that performs best globally
- Qualitative analysis points to leakage between R and A
- Switching order of R and A improves results for Reasoning (Q → AR)

Model	Question Answering[1]	Answer Justification[1]	Reasoning
R2C[2]	52.3	**61.8**	33.2
Ours	**89.5**	41.2	**36.1**
Human	91.0	93.0	85.0

[1]Results not directly comparable

[2]Recognition to Cognition Networks from Zellers et al. (CVPR 2019)

VQA via Reason Prediction

Mihir Rana and Kenil Tanna

Ablation Results on VCRSmall-Val

- BERT (text-only) does extremely well on answering tasks
- Visual features not very important
- Given the reason, question not very important (probably due to leakage)

Model	QR → A	R → A	Q → R	Q → AR
BERT	**89.6**	**85.9**	38.1	34.2
No Vision	89.3	85.5	39.5	35.5
Full	89.3	85.5	**40.4**	**36.0**

Building a Semantic Parser Over a Very Long Period of Time

Method: paraphrase model with domain-specific grammar

Natural Language Query

How many songs were released by Taylor Swift in 2014?

Canonical Utterance

number of songs whose artist is Taylor Swift and whose year is 2014

Lambda DCS Logical Form

$\text{count}(\mathbf{R}(\text{songs})).(\text{artist.TaylorSwift} \sqcap \text{year.2017})$

Answer

16

Main Results and Experiments with Sample Size

model	accuracy	oracle accuracy
baseline	50.3	68.6
final	65.4	76.5

Table: Main results

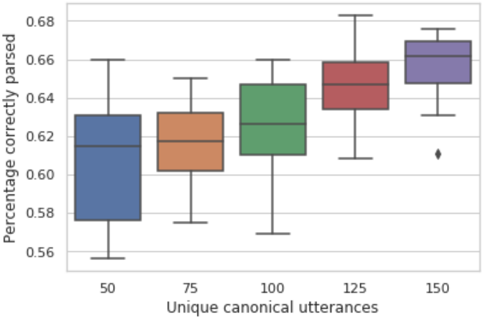


Figure: Percent of phrases parsed correctly for different size of the training data

Error Analysis and Conclusions

Error Analysis

- Numeric queries were often incorrectly parsed
  "What songs are by more than two artists?"
- Queries of types not in training data were often incorrectly parsed
  "What songs came out after 2006?"

Conclusions

- Types of queries demonstrated by the training data matter
- Total size of trianing data does not matter
- Flexibility remains an issue which neural nets may address

A Transfer -Learning Approach to Detect Duplicate Questions in Stack Exchange

Xiaoyi Zhang
Daoyang Shan
Yihong Zhou
Ziwei Wang

NYU Center for Data Science

Facts:

1. Manual labeling -> bad user experience
2. Previous attempts give precision unsuitable (~60%) for industrial application.

Challenges & what is known:

1. Common issues of user-generated context.
2. Specific to Stack Exchange: slight semantic mismatch can refer implicitly to the same answer.
   (e.g. **Am I Jewish? V.S. Does a Jewish grandmother get one accepted as Jew?**)
3. Feature extraction, bag-of-words, TF-IDF, ConvNet do NOT give accurate results.

Question:

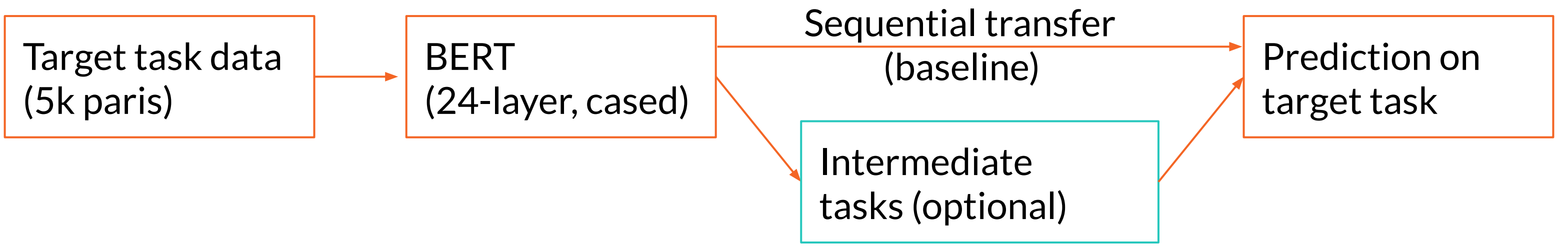
Is it ever possible to build a framework that well-captures the semantic patterns of duplicate questions while does not take too long to train?
YES! By transfer learning from a large pre-trained language model.

Methodology

Dataset:

- **Source**: Stack Exchange data dump (since its launch), top-100 popular subforums, in English.
- **Quantity**: 250k questions pairs, among which 100k marked as duplicate by admin.

Framework:



Training time: 50 - 80 min on NVIDIA TESLA-p40 GPU

Results

Metric	*Sequential*	*CoLA*	*MRPC*	*STS-B*	*0.3 STEX*	*0.6 STEX*	*0.9 STEX*
Accuracy_F1	0.904	0.885	0.870	0.892	0.894	0.905	0.909
Accuracy	0.916	0.904	0.887	0.908	0.908	0.915	0.919
F1	0.893	0.865	0.853	0.875	0.880	0.894	0.899
Precision	0.921	0.919	0.840	0.901	0.866	0.909	0.913
Recall	0.867	0.818	0.867	0.851	0.895	0.880	0.884

1. Intermediate tasks do not guarantee better performance
2. Higher ratio of STEX in intermediate tasks enhances overall precision
3. Baseline comparable to human judgement, and certain characters of questions pairs can either favor or confuse the model.
   i. Model beats Humans: (TRUE: 1, Model: 1, Manual: 0)
   ➢ **Am I Jewish? V.S. Does a Jewish grandmother get one accepted as Jew?**
   ii. Humans beat Model: (TRUE: 0, Model: 1, Manual: 0)
   ➢ **Voltage divider to measure battery voltage on Arduino V.S. Solving differential equations numerically using Arduino**

SOTU-TIME: A Scheme and Corpus for Classifying Temporal Orientation in Political Speech

- “I am going to run for President.”
- “Yes we can!”
- “Make America great again!”
- “We will always honor their memory.”

1. Reflection on the data

- „We must strengthen the economy.“

3. Scheme

	Past	Future
TRUE		
FALSE		

3. Annotation Guidelines

4. Annotation

- Sample of 3000 sentences

Experimental Results

Models:

- Support Vector Machines
- Bidirectional Gated Recurrent Units
- Stacked BiLSTM

Representations:

- Bag of words, POS tags, Bigrams
- Word embeddings and POS embeddings

Baseline(based on Rules):
Past Accuracy = 72%
Future Accuracy = 56.94%

Model	Accuracy (%)	P | R	F1
SVM(BOW)	73.01	0.74 | 0.73	0.72
SVM(BOW+POS)	75.54	0.76 | 0.76	0.75
SVM(BOW+POS+ Bigrams)	78.4	0.79 | 0.78	0.78
RNN (W2V)	79.08	0.88 | 0.63	0.73
RNN with attention (W2V)	**79.43**	0.89 | 0.62	0.73
LSTM (GloVe)	78.75	0.79 | 0.79	**0.79**
LSTM (with Word and POS embeddings)	72.34	0.78 | 0.72	0.70

Past Orientation Task

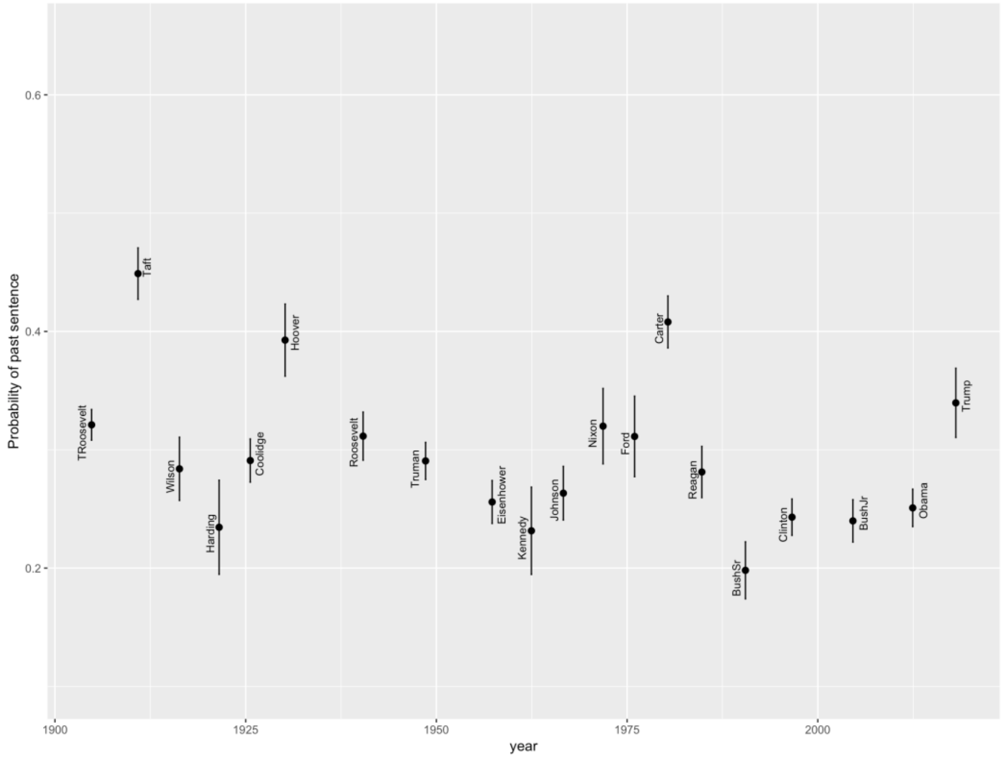
Model	Accuracy (%)	P | R	F1
SVM(BOW)	75.7	0.76 | 0.76	0.76
SVM(BOW+POS)	75.05	0.76 | 0.76	0.76
SVM(BOW+POS+ Bigrams)	80.27	0.80 | 0.80	0.80
RNN (W2V)	83.47	0.84|0.84	0.836
RNN with attention (W2V)	**83.81**	0.83 | 0.85	**0.842**
LSTM (GloVe)	82.8	0.83 | 0.83	0.83
LSTM (with Word and POS embeddings)	76.22	0.76 | 0.76	0.76

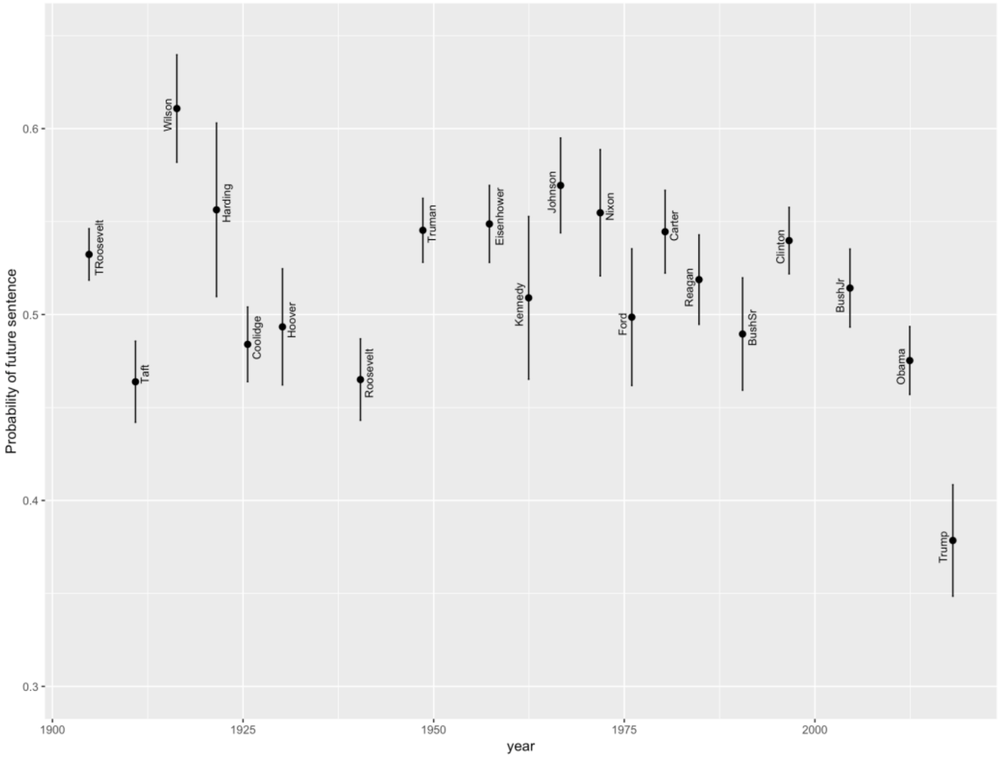
Future Orientation Task

Past & Future references in the SOTU

Past

Future

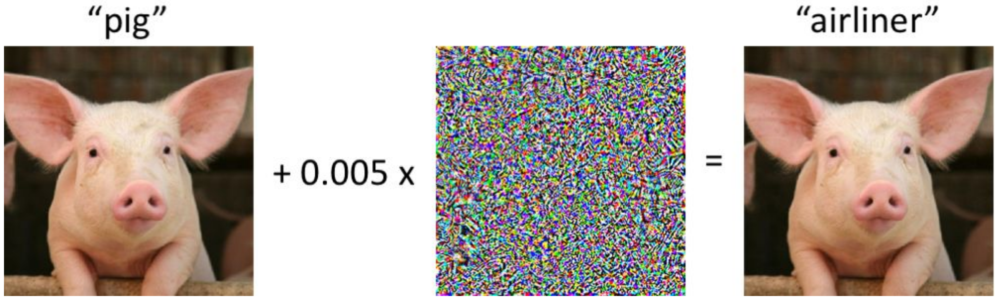




Team 24: Oliver Aastrand, Kelly Marshall, Chiao-Hsun Wang

Idea

-Models often don’t learn what we want them to. Adversarial examples take advantage of a model’s weaknesses to “break” its performance



-We apply this to Natural Language Inference

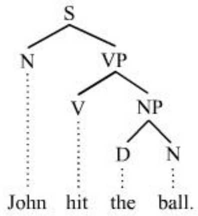
Approach: Negation

Example : A dog is in the water

Two ways of negation:

1. **Negate the verbs** : A dog is not in the water
2. **It is not the case that** : It is not the case that a dog is in the water

Logical rules: **ent(s1, s2) -> ent(~s2 , ~s1)**



Constituency-based parse tree

A wet brown dog swims **entails** A dog is in the water.

A dog is not in the water **entails** A wet brown dog does not swim.

Result

Negation done via parse trees:

Model : MT-DNN (SOTA)

Accuracy on SNLI dataset: **90.67%**

ent(s1,s2) -> ent(~s2, ~s1) + negation of verbs: **22.47%**

ent(s1,s2) -> ent(~s2, ~s1) + “It is not the case”: **6.53%**

Motivation: Formal vs. Informal Text

- Summarization is traditionally done on formal text
  - Formal text: usually by strict guidelines
    - news, journal articles, academic papers, etc.
  - Informal text: personal, casual, abundant, slangish
    - emails, text messages, tweets, etc.
    - Leads to practical implementations

- Enron Corpus: public, uncensored, natural
  - Sent (first) emails only

Text Summarization for Email Subject Lines

Model & Results

Encoder-Decoder with Attention

Abstractive

inspired by machine translation

- (Rush et al. 2015, Bahdanau et al. 2014)

TextRank

Extractive

Graph-based Model

	Formal		Informal	
	DUC-2004	Gigaword	Enron (ABS)	Enron (TextRank)
Rouge-1	28.18	31.00	19.43	17.06
Rouge-2	8.49	12.65	10.54	2.41
Rouge-L	23.81	28.34	21.22	18.55

Text Summarization for Email Subject Lines

Result Analysis

- Attention based model captured the key idea of the email
  - Lunch, friday, april 13
- "FREE" is in the title but not in the body of this email
  - "will be provided" = "FREE" ?
- Requires much deeper understanding of the semantics

Example: Extractive vs. Abstractive

Email:

Thanks for all your hard work and happy birthday! Lunch will be provided on friday, April 13, by Tim Belden and Chris Calger to everyone on the floor as a thanks for all you 've done for enron this month. We 'll also celebrate this month 's birthdays by having cookies for everyone .

Subject:

FREE LUNCH on Friday, April 13

TextRank (Extractive):

We'll also celebrate this month's birthdays by having cookies for everyone.

Attention (Abstractive):

lunch friday, april 13

Email Classification

Problem Introduction

Inbox 104

Starred

Snoozed

Sent

Johann Brehmer 2 [DS General] CDS Seminar // 2019-05-01 // Richard Shen (Wayve): End-to-end policy learn...

NYU Office of Alumn. Enjoy the Great Weather with a Great Time—Join Your NYU Friends at our May Events - Vie... events

Message-ID: <26804150.1075842955435.JavaMail.evans@thyme>
Date: Tue, 29 Aug 2000 09:11:00 -0700 (PDT)
From: jeff.dasovich@enron.com
To: gramlr@pjm.com
Subject: Re: FW: Possible co-sponsorships

Though I had a somewhat different notion when I initially raised the
idea of
co-sponsorship, I agree with Lee's observations and think that we
should
proceed the way he suggests.

Lee

predict

- Company Business, Strategy
- Purely Personal
- Personal but in Professional Context
- Logistic Arrangements
- Employment Arrangements
- Document Editing/Checking

4.43 GB (29%) of 15 GB used
Manage

Terms · Privacy · Program Policies

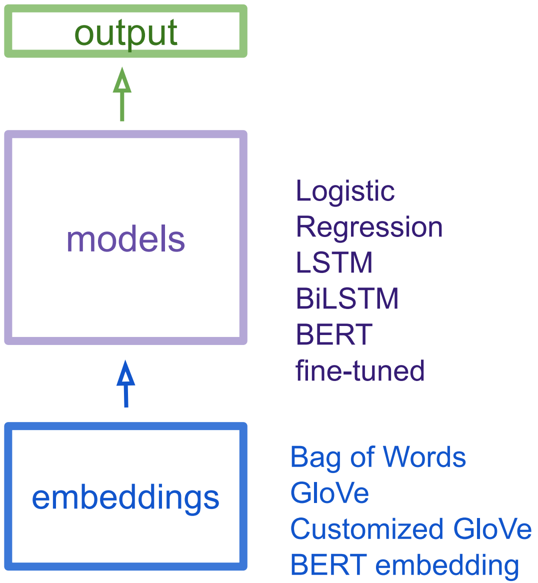
Last account activity: 26 minutes ago
Details

Email Classification

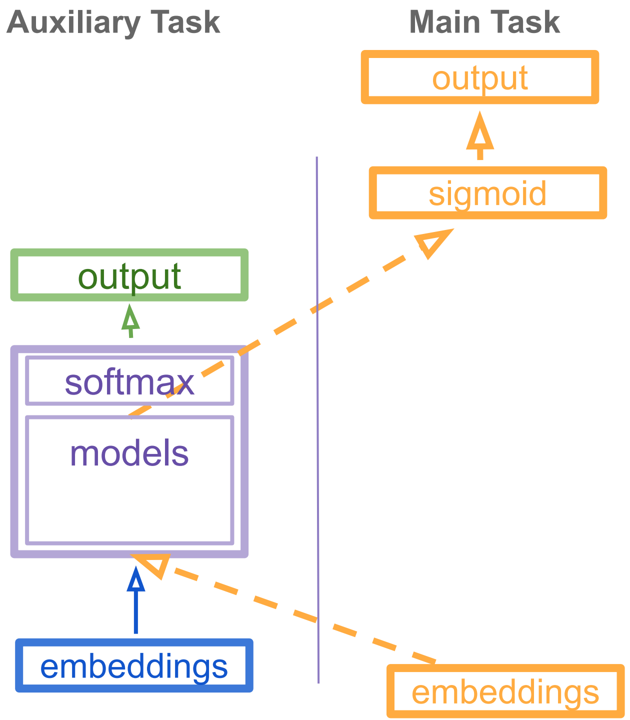
Model Architecture

- Inbox 104
- Starred
- Snoozed
- Sent
- Drafts 1
- Notes
- Other
  - Coursera
  - Media Post
  - peerstream
- Less

General Model Architecture



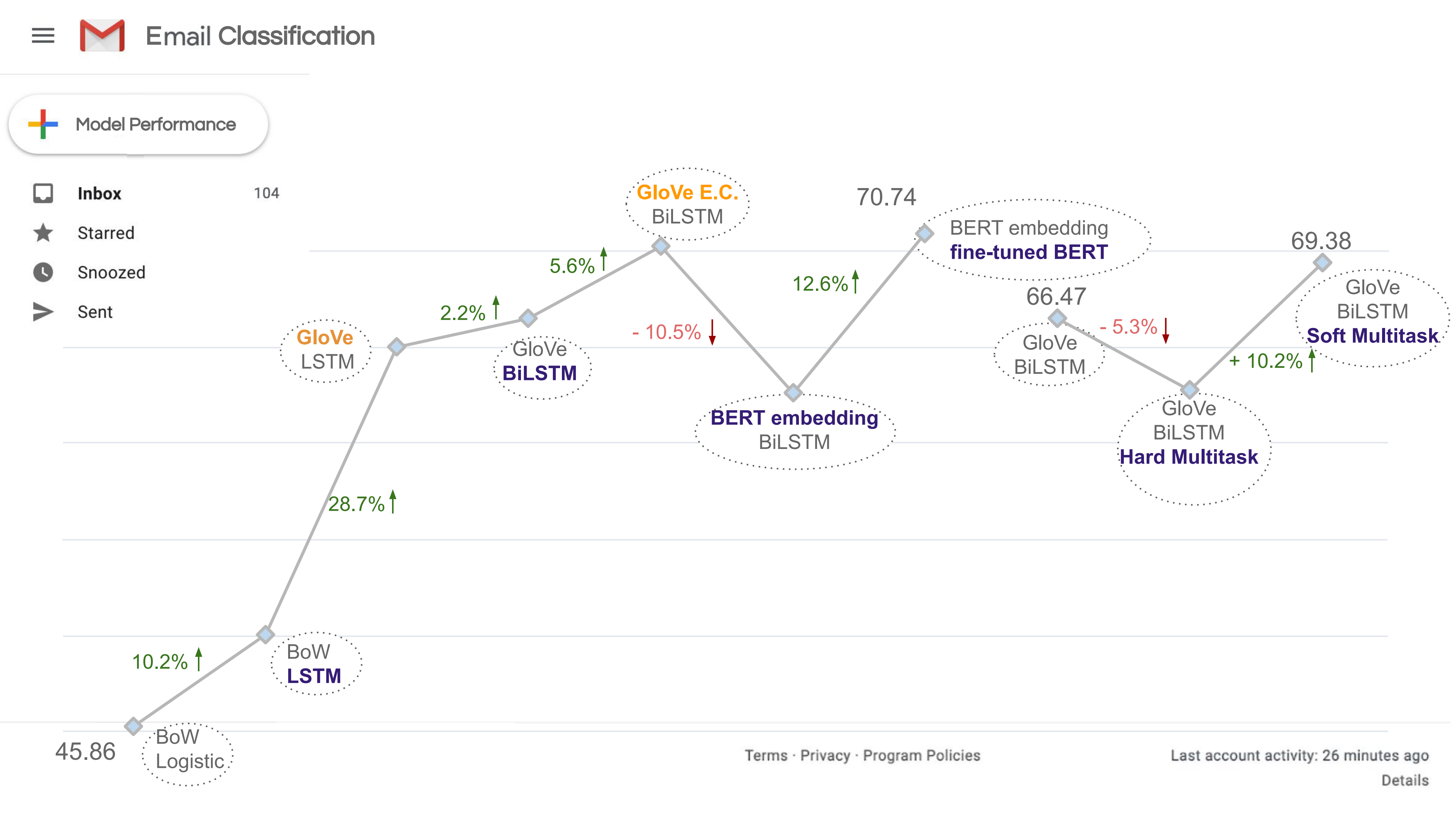
Multi-task Architecture



4.43 GB (29%) of 15 GB used
Manage

Terms · Privacy · Program Policies

Last account activity: 26 minutes ago
Details

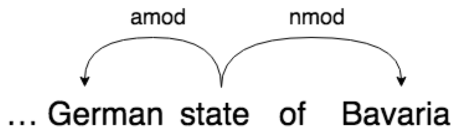

Email Classification
Model Performance
Inbox 104
Starred
Snoozed
Sent
45.86
BoW
Logistic
10.2%
BoW
LSTM
28.7%
GloVe
LSTM
2.2%
GloVe
BiLSTM
5.6%
GloVe E.C.
BiLSTM
- 10.5%
BERT embedding
BiLSTM
12.6%
70.74
BERT embedding
fine-tuned BERT
66.47
GloVe
BiLSTM
- 5.3%
GloVe
BiLSTM
Hard Multitask
+ 10.2%
69.38
GloVe
BiLSTM
Soft Multitask
Terms · Privacy · Program Policies
Last account activity: 26 minutes ago
Details

Dependency-Enhanced Attention for Fact Verification

Claim: Munich is the capital of Germany.
Retrieved Evidence:
[wiki/Germany]
Germany's capital and largest metropolis is Berlin, while its largest conurbation is the Ruhr (main centres: Dortmund and Essen).

[wiki/Munich]
Munich is the capital and largest city of the German state of Bavaria, on the banks of River Isar north of the Bavarian Alps. Following a final reunification of the Wittelsbachian Duchy of Bavaria, previously divided and sub-divided for more than 200 years, the town became the country's sole capital in 1506. Having evolved from a duchy's capital into that of an electorate (1623), and later a sovereign kingdom (1806), Munich has been a major European centre of arts, architecture, culture and science since the early 19th century, heavily sponsored by the Bavarian monarchs.

Label: Support



Motivation:
Existing models often fail to extract precise relationships among words in long, complicated sentences

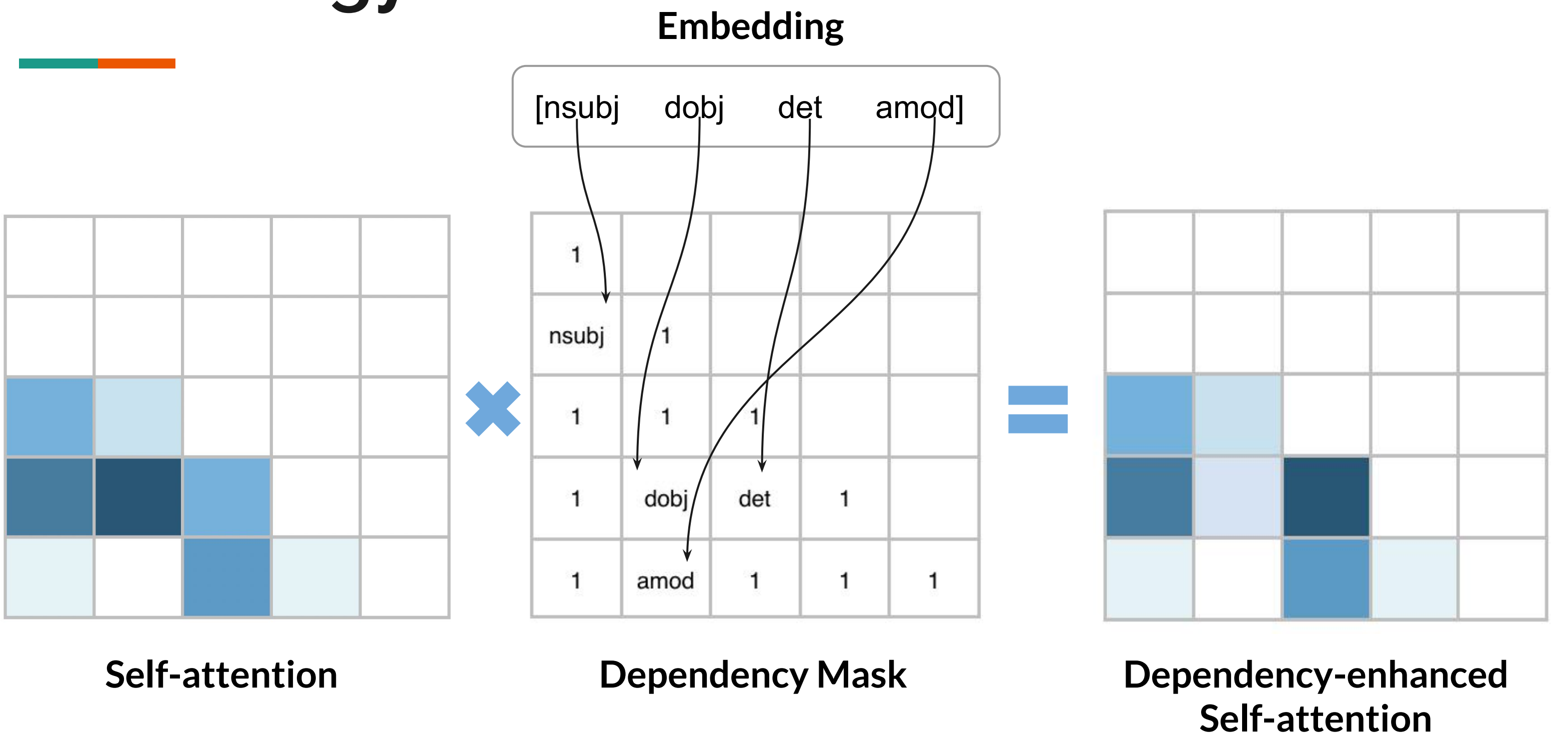
Our Task:
- Learn an embedding for dependency types
- Use dependency-enhanced self-attention in NLI

Goal:
Improve the understanding of relationships among words in complex sentences

Sample from FEVER dataset

Group 27: Nimi Wang, Fangjun Zhang, Ruoyu Zhu

Methodology



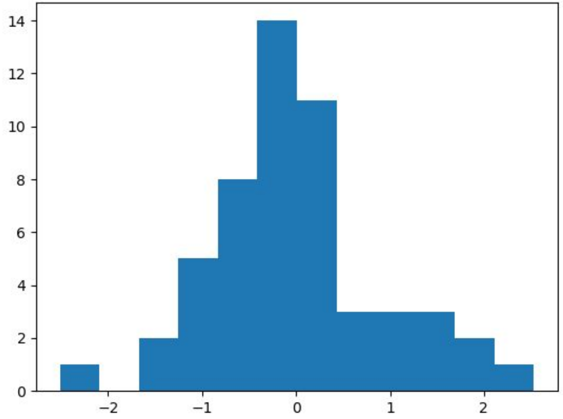
Experiment and Result

Use a subset (~10k) of FEVER data for training

Quantitative Results

	Label Accuracy	FEVER Score
ESIM	0.743	0.685
ESIM + Self-Attention	0.739	0.682
ESIM + Dep.-Enhanced Self-Attn.	**0.744**	**0.689**

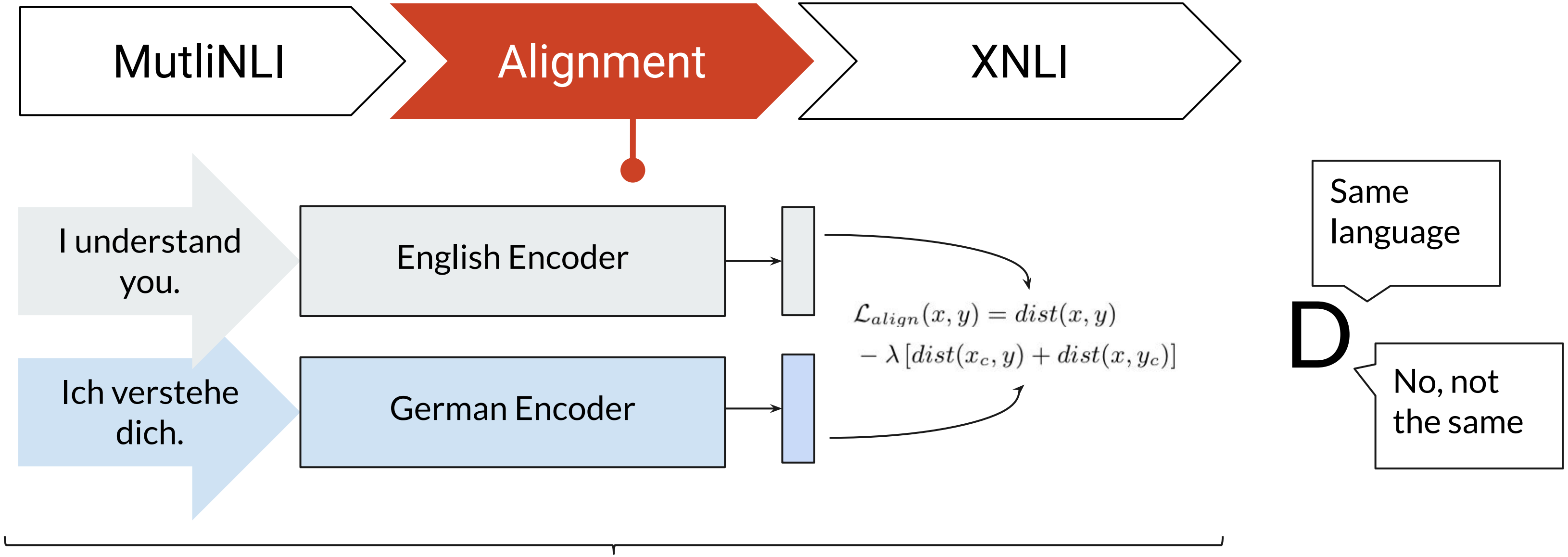
Histogram of Dependency Embedding



The Task: Learning Cross-Lingual Sentence Representations for Natural Language Inference

Asena D. Cengiz, Gauri Sarode, & Samantha Petter

The Approach



XNLI (Conneau et al, 2018)

Results

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
XNLI Baselines from (Conneau et al., 2018)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	**67.7**	**68.7**	**67.7**	**68.9**	**67.9**	**65.4**	**64.2**	**64.8**	**66.4**	**64.1**	**65.8**	**64.1**	**55.7**	**58.4**
Baseline Results: *XNLI Multilingual Sentence Encoder (Our Implementation - Test Acc %)*															
X-BiLSTM-max	**70.1**	64.6	62.1	61.2	59.3	58.2	58.7	59.4	58.1	58.5	56.3	55.7	56.2	-	-
Model Results: *XNLI Multilingual Sentence Encoder + Discriminator (Test Acc %)*															
X-BiLSTM-max	**70.1**	65.0	63.7	61.5	59.7	58.1	59.0	58.3	58.6	60.0	57.1	56.6	55.9	-	-

Table 1: Cross-lingual natural language inference (XNLI) test accuracy for 13 languages.