Deep Neural Net Approaches for Natural Language Processing

Gary Geunbae Lee
Dept of CSE/AIGS
POSTECH
A few applications of HLT/NLP/CL

- Spelling correction, grammar checking
  ...(language learning and evaluation e.g. TOEFL essay score)
- Better search engines
- Information extraction, gisting
- Psychotherapy; chat bot, etc.
- Speech recognition (and text-to-speech)
- Dialogue systems
- Machine translation; speech translation
- Trans-lingual summarization, detection, extraction ...
Question Answering: IBM’s Watson

Won Jeopardy on February 16, 2011!

WILLIAM WILKINSON’S “AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA” INSPIRED THIS AUTHOR’S MOST FAMOUS NOVEL

Bram Stoker
Hi Dan, we’ve now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris
Information Extraction & Sentiment Analysis

Attributes:
- zoom
- affordability
- size and weight
- flash
- ease of use

Size and weight

- nice and compact to carry!
- since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- the camera feels flimsy, is plastic and very light in weight you have to be very delicate in the handling of this camera
Machine Translation

- Helping human translators

Fully automatic

Enter Source Text:

这不过是一个时间的问题.

Translation from Stanford’s Phrasal:

This is only a matter of time.
Language Technology

Spam detection
Let’s go to Agra!
Buy V1AGRA ...
✓
✗

Part-of-speech (POS) tagging
ADJ  ADJ  NOUN  VERB  ADV
Colorless  green  ideas  sleep furiously.

Named entity recognition (NER)
PERSON  ORG  LOC
Einstein met with UN officials in Princeton

Sentiment analysis
Best roast chicken in San Francisco!
The waiter ignored us for 20 minutes.

Coreference resolution
Carter told Mubarak he shouldn’t run again.

Word sense disambiguation (WSD)
I need new batteries for my mouse.

Parsing
I can see Alcatraz from the window!

Machine translation (MT)
第13届上海国际电影节开幕...
The 13th Shanghai International Film Festival...

Information extraction (IE)
You’re invited to our dinner party, Friday May 27 at 8:30

Question answering (QA)
Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
XYZ acquired ABC yesterday
ABC has been taken over by XYZ

Summarization
The Dow Jones is up
The S&P500 jumped
Economy is good
Housing prices rose

Dialog
Where is Citizen Kane playing in SF?
Castro Theatre at 7:30. Do you want a ticket?
Levels of Language

- **Phonetics/phonology/morphology:** what words (or subwords) are we dealing with?
- **Syntax:** What phrases are we dealing with? Which words modify one another?
- **Semantics:** What’s the literal meaning?
- **Pragmatics:** What should you conclude from the fact that I said something? How should you react?
What’s hard - ambiguities, ambiguities, all different levels of ambiguities

John stopped at the donut store on his way home from work. He thought a coffee was good every few hours. But it turned out to be too expensive there. [from J. Eisner]

- donut: To get a donut (doughnut; spare tire) for his car?
- Donut store: store where donuts shop? or is run by donuts? or looks like a big donut? or made of donut?
- From work: Well, actually, he stopped there from hunger and exhaustion, not just from work.
- Every few hours: That’s how often he thought it? Or that’s for coffee?
- it: the particular coffee that was good every few hours? the donut store? the situation
- Too expensive: too expensive for what? what are we supposed to conclude about what John did?
Why else is natural language understanding difficult?

- **non-standard English**
  
  Great job @justinbieber! Were SOO PROUD of what you’ve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

- **segmentation issues**
  
  the New York-New Haven Railroad

- **idioms**
  
  dark horse
  get cold feet
  lose face
  throw in the towel

- **neologisms**
  
  unfriend
  Retweet
  bromance

- **world knowledge**
  
  Mary and Sue are sisters.
  Mary and Sue are mothers.

- **tricky entity names**
  
  Where is *A Bug’s Life* playing ...
  *Let It Be* was recorded ...
  ... a mutation on the *for* gene ...

But that’s what makes it fun!
Imagine:

- Each sentence $W = \{ w_1, w_2, ..., w_n \}$ gets a probability $P(W|X)$ in a context $X$ (think of it in the intuitive sense for now).

- For every possible context $X$, sort all the imaginable sentences $W$ according to $P(W|X)$:
  - Ideal situation:
    - best sentence (most probable in context $X$)
    - NB: same for interpretation

```
P(W)
```

```
“ungrammatical” sentences
```
Artificial Neural Networks (ANN)

Axon Terminal Branches of Axon

Activation Function

Dendrites

Terminal Branches of Axon
Layered Networks

Output:

\[ y = f(\sum w_i x_i) \]
Deep learning Innovation

- Combining Feature Learning and Classification as Unified Framework  ※ Learning what to learn, how to learn

Feature learning aspect of DNN based Image Classification
Vanilla recurrent neural networks (RNNs)

- RNNs have connections from the outputs of previous time steps to inputs of next time steps.

- For sequential data, a RNN usually computes hidden state $h_t$ from the previous hidden state $h_{t-1}$ and the input $x_t$.
  - $h_t = \sigma(W_h h_{t-1} + W_x x_t + b)$
Vanishing gradient problem

- \( h_t = \sigma(W_h h_{t-1} + W_x x_t + b) \)
- Let’s assume \( \sigma \) is the identity function

\[
\frac{\partial J^{(i)}(\theta)}{\partial h^{(j)}} = \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} \prod_{j < t \leq i} \frac{\partial h^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} \prod_{j < t \leq i} W_h = \frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} W_h^\ell
\]

If all \( \frac{\partial h^t}{\partial h^{t-1}} < 1 \) \( \Rightarrow \frac{\partial J^t}{\partial h^1} \approx 0 \)
Long short-term memory networks (LSTMs)

- LSTMs explicitly keep and update cell memory $c^{(t)}$ by
  - Removing the previous cell content $c^{(t-1)}$ by multiplying it with $f^{(t)}$
  - Adding the new cell content $\tilde{c}^{(t)}$ multiplied by $i^{(t)}$
- LSTMs produce output $h^{(t)} = o^{(t)} \circ \text{tanh} c^{(t)}$

\[
\begin{align*}
 f^{(t)} &= \sigma \left( W_f h^{(t-1)} + U_f x^{(t)} + b_f \right) \\
 i^{(t)} &= \sigma \left( W_i h^{(t-1)} + U_i x^{(t)} + b_i \right) \\
 o^{(t)} &= \sigma \left( W_o h^{(t-1)} + U_o x^{(t)} + b_o \right) \\
 \tilde{c}^{(t)} &= \text{tanh} \left( W_c h^{(t-1)} + U_c x^{(t)} + b_c \right) \\
 c^{(t)} &= f^{(t)} \circ c^{(t-1)} + i^{(t)} \circ \tilde{c}^{(t)} \\
 h^{(t)} &= o^{(t)} \circ \text{tanh} c^{(t)}
\end{align*}
\]
Gated recurrent units (GRUs)

• GRUs keeps and update $h^{(t)}$ by two gates:
  • Update gate $u^{(t)}$ decides
    • How much the old hidden representation $h^{(t)}$ is removed
    • how much the new hidden representation $\tilde{h}^{(t)}$ is added
  • Reset gate $r^{(t)}$ decides how much old representation $h^{(t)}$ is needed
to compute new representation $\tilde{h}^{(t)}$

• GRUs also use less number of gates and have smaller parameters than LSTMs

\[
\begin{align*}
  u^{(t)} &= \sigma \left( W_u h^{(t-1)} + U_u x^{(t)} + b_u \right) \\
  r^{(t)} &= \sigma \left( W_r h^{(t-1)} + U_r x^{(t)} + b_r \right) \\
  \tilde{h}^{(t)} &= \tanh \left( W_h (r^{(t)} \circ h^{(t-1)}) + U_h x^{(t)} + b_h \right) \\
  h^{(t)} &= (1 - u^{(t)}) \circ h^{(t-1)} + u^{(t)} \circ \tilde{h}^{(t)}
\end{align*}
\]
Bidirectional Multi-Layer RNNs

the movie was terribly exciting!

19
Parallel computing for Deep Learning

- History of parallel/distributed systems for Deep Learning computing

- Google taps 16k computers to look for cats –for Science!

- Univ. of Toronto uses 2 GPUs for 1.2M training images for 1000 classes image classification (※ ImageNet Large Scale Visual Recognition Challenge)

- Stanford uses 12 GPUs for Large-scale Video Classification with Convolutional Neural Networks (※ 10M Youtube video)

- Google uses 16K CPU cores for training 22-layers Deep neural network (※ GoogLeNet, 2014)

- Baidu’s Artificial Intelligence Supercalculator Beats Google at Image Recognition
Deep Learning for NLP
Word Vector

- Represent words as vectors

\[
\text{expect} = \begin{pmatrix}
0.286 \\
0.792 \\
-0.177 \\
-0.107 \\
0.109 \\
-0.542 \\
0.349 \\
0.271 \\
0.487
\end{pmatrix}
\]
Word Vector

• Distributional semantics: A word’s meaning is given by the words that frequently appear close-by

• “You shall know a word by the company it keeps”

• Word2vec objective function (skip-grams)

\[
J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{-m \leq j \leq m \atop j \neq 0} \log P(w_{t+j} | w_t; \theta)
\]
A word’s **contextual embedding** must consider its context.
ELMo: Embeddings from Language Model

- Multi-layer bidirectional LSTM language model

\[ R_k = \{ x_k^{LM}, \overrightarrow{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM} \mid j = 1, \ldots, L\} \]
\[ = \{ h_{k,j}^{LM} \mid j = 0, \ldots, L\}, \]
\[ h_{k,0}^{LM} = x_k^{LM} \quad \text{(token representation; GloVe)} \]
\[ h_{k,j}^{LM} = \overrightarrow{h}_{k,j}^{LM}, \overleftarrow{h}_{k,j}^{LM} \quad \text{(LSTM state)} \]

\[ \text{ELMo}_k^{\text{task}} = E(R_k; \Theta^{\text{task}}) = \gamma^{\text{task}} \sum_{j=0}^{L} s_j^{\text{task}} h_{k,j}^{LM} \]

\( \gamma^{\text{task}} \): scale (hyper-parameter)
\( s_j^{\text{task}} \): weight (learned)
ELMo for MRC

- ELMo as a word embedding
Transformer

- Parallel self-attention
  - Looks at self, and determines where to focus

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]

Q,K,V – vectors for every word, output attention summed up;
head means different Q,K,V vector with different weights

Vaswani, Ashish, et al. "Attention is all you need." NIPS 2017
• Training 1. Masked words prediction
  • 15% of words are [MASK]ed

*GELU: Gaussian error linear unit
BERT: Bidirectional Encoder Representations from Transformers

• Training 2. Next sentence prediction
  • To understand texts more than a sentence

Input = [CLS] the man went to [MASK] store [SEP]
  he bought a gallon [MASK] milk [SEP]
Label = IsNext

Input = [CLS] the man [MASK] to the store [SEP]
  penguin [MASK] are flight ##less birds [SEP]
Label = NotNext
BERT: Bidirectional Encoder Representations from Transformers

- BERT as universal pre-trained model for NLP
  - BERT requires minimal additional layers and fine-tuning

*GLUE benchmark task

In pre-training, optimize $L_1(u)$

$u$: Unlabeled dataset

$\Theta$: Model parameters

In fine-tuning, optimize $L_3(c)$

$c$: Labeled dataset

$\lambda$: Hyper-parameter weight

$*GPT-2/3$: zero/few-shot learning
Sequence labeling

• Sequence labeling is the task of assigning a categorical label to each member of an observed sequence.

• Examples of sequence labeling
  
  • **Part-of-speech tagging** labels each word with a grammatical category.
    
    • e.g. The | trees | are | ... → DT (determiner) | NNS (plural noun) | VBP (plural verb)
  
  • **Named entity recognition** locates and classifies named entity in text. It can be tackled by labeling each word with a named entity category.

    • e.g Barack | Obama | said | ... → B-PERSON | I-PERSON | O | ...
Bi-directional LSTM-CNNs-CRF

• **Bi-directional LSTMs** encode word embeddings and character representations

• **Conditional random fields** compute the distribution of output sequence
  • Viterbi algorithm is applied during training and decoding
  • The objective is the negative log-likelihood of the output sequence distribution

\[
p(y|z; W, b) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, z)}{\sum_{y' \in Y(z)} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_i, z)}
\]

• \( z \): input sequence
• \( y \): output sequence
• \( Y(z) \): a set of all possible output sequences when given the input sequence \( z \)

[Ma 2016]
Seq2Seq NMT via fixed-length representations

- Encoder RNN compresses input sequence into a fixed-length representation
- Decoder RNN produces output sequence from the representation
  - Each produced output token is fed into the next RNN’s input

[Sutskever 2014]
S2S NMT with attention mechanism

• It’s hard to encode all the information of an input sequence into a fixed-length representation.

• We can focus important parts of input sequences for each decoding step by attention mechanism.

[Bahdanau 2015]
MNMT (Multilingual Neural Machine Translation)

“Multilingual Neural Machine Translation with Knowledge Distillation”
(Microsoft Research, ICLR 2019)

- **Knowledge Distillation**: Use the knowledge of big & well-trained networks when training small model

Teacher Models (Expert) → Knowledge Distillation → Multilingual Student model
"Multilingual Neural Machine Translation with Knowledge Distillation"

- Applying Knowledge Distillation by adjusting Loss Function

- **Loss Function**
  
  \[
  \mathcal{L}_{\text{ALL}}(D; \theta, \theta_T) = (1 - \lambda)\mathcal{L}_{\text{NLL}}(D; \theta) + \lambda\mathcal{L}_{\text{KD}}(D; \theta, \theta_T),
  \]

  *Selective Distillation*: only when "Individual accuracy > Multilingual accuracy" apply Lkd loss, else apply LNLL loss

- **Log-likelihood of (MNMT)**

  \[
  \mathcal{L}_{\text{NLL}}(D; \theta) = -\sum_{(x,y) \in D} \log P(y|x; \theta),
  \]

  \[
  \log P(y|x; \theta) = \sum_{t=1}^{T_y} \sum_{k=1}^{|
  V|} \mathbb{1}\{y_t = k\} \log P(y_t = k|y_{<t}, x; \theta),
  \]

- **Log-likelihood of (MNMT with KD)**

  \[
  \mathcal{L}_{\text{KD}}(D; \theta, \theta_T) = -\sum_{(x,y) \in D} \sum_{t=1}^{T_y} \sum_{k=1}^{|
  V|} Q\{y_t = k|y_{<t}, x; \theta_T\} \log P(y_t = k|y_{<t}, x; \theta).
  \]

  Teacher’s output distribution

*Better than interpolation*
MNMT (Multilingual Neural Machine Translation)

- Based on Transformer Encoder-Decoder Model

```
Ko → En
Ja → En
Vi → En
Zh → En
```

![Diagram of MNMT model]

- Teacher model
  - Ko-En topk probability
  - Ja-En topk probability
  - Vi-En topk probability
  - Zh-En topk probability

- Student model
  - Ko → En
  - Vi → En

*KD

- Add & Norm
- Position-wise FFNN
- Multi-head Attention
- Masked Multi-head Self-Attention
- Positional Encoding
- embedding

- num_layers ×

- softmax
- Dense

- num_layers ×

- Add & Norm
- Position-wise FFNN
- Multi-head Attention
- Masked Multi-head Self-Attention
- Positional Encoding
- embedding
Dependency parsing

• Dependency parsing is the task of extracting **dependencies** between **head** and **dependent** words from a sentence.

```
ROOT  He     has   good   control .
    PRP    VBZ    JJ    NN   
```

• A dependency is the arrow from a head to a dependent with a grammatical type called **relation** (e.g. nsubj).

• Dependencies show which words depend on (modify or are arguments of) which other words.
Neural transition-based dependency parsing

- Extract features from Stack and Buffer
  - lc/rc: leftmost/rightmost children
- Classify an action by neural networks
  - The objective is the negative log-likelihood of the action distribution

Feature extraction

Feedforward neural network-based action classifier

[Chen 2014]
Semantic parsing (weakly supervised)

- Semantic parsing is a task of mapping **natural language** to programs
- We aim to develop semantic parsers **without direct supervision on programs**

```
Program
(map (argmax (filter all-rows (λ (x) (= (string:country x) "greece")))
  index)
number:year)
```

```
Natural language
“Greece held its last Summer Olympics in which year?”
```

```
Context
<table>
<thead>
<tr>
<th>Year</th>
<th>City</th>
<th>Country</th>
<th>Nations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1896</td>
<td>Athens</td>
<td>Greece</td>
<td>14</td>
</tr>
<tr>
<td>1900</td>
<td>Paris</td>
<td>France</td>
<td>24</td>
</tr>
<tr>
<td>1904</td>
<td>St. Louis</td>
<td>USA</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>2004</td>
<td>Athens</td>
<td>Greece</td>
<td>201</td>
</tr>
<tr>
<td>2008</td>
<td>Beijing</td>
<td>China</td>
<td>204</td>
</tr>
<tr>
<td>2012</td>
<td>London</td>
<td>UK</td>
<td>204</td>
</tr>
</tbody>
</table>
```

```
Denotation 2004
```
• Bottom-up neural semantic parsing
  • Neural networks decodes logic actions
  • Grammar of logical form is restricted (function definition)

\[
\begin{align*}
  (\text{Hop } r \ p) & \Rightarrow \{ e_2 | e_1 \in r, (e_1, p, e_2) \in K \} \\
  (\text{ArgMax } r \ p) & \Rightarrow \{ e_1 | e_1 \in r, \exists e_2 \in E : (e_1, p, e_2) \in K, \forall e : (e_1, p, e) \in K, e_2 \geq e \} \\
  (\text{ArgMin } r \ p) & \Rightarrow \{ e_1 | e_1 \in r, \exists e_2 \in E : (e_1, p, e_2) \in K, \forall e : (e_1, p, e) \in K, e_2 \leq e \} \\
  (\text{Filter } r_1 \ r_2 \ p) & \Rightarrow \{ e_1 | e_1 \in r_1, \exists e_2 \in r_2 : (e_1, p, e_2) \in K \}
\end{align*}
\]

Table 1: Interpreter functions. \( r \) represents a variable, \( p \) a property in Freebase. \( \geq \) and \( \leq \) are defined on numbers and dates.

![Diagram](https://via.placeholder.com/150)

*freebase entity, property matching
*off-line generation of logic Program (sampling)

Figure 2: Semantic Parsing with NSM. The key embeddings of the key-variable memory are the output of the sequence model at certain encoding or decoding steps. For illustration purposes, we also show the values of the variables in parentheses, but the sequence model never sees these values, and only references them with the name of the variable ("R1"). A special token "GO" indicates the start of decoding, and "Return" indicates the end of decoding.

Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision, Liang et al., ACL 2017

Memory Augmented Policy Optimization for Program Synthesis and Semantic Parsing, Liang et al., NIPS 2018
Top-down neural semantic parsing
  - Neural net generate derivation actions
  - Type system reduce search space (entity matching)

Figure 1: Overview of our semantic parsing model. The encoder performs entity embedding and linking before encoding the question with a bidirectional LSTM. The decoder predicts a sequence of grammar rules that generate a well-typed logical form.

Neural Semantic Parsing with Type Constraints for Semi-Structured Tables, Krishnamurthy et al., EMNLP 2017
Iterative Search for Weakly Supervised Semantic Parsing, Dasigi et al., NAACL 2019
**Abstract Syntax Description Language (ASDL)** for Python

Code: `sorted(my_list, reverse=True)`

*parse tree to verify python code

**ASDL for SQL**

```python
stmt = Select(agg_op? agg, idx column_idx, cond_expr* conditions)
cond_expr = Condition(cmp_op op, idx column_idx, string value)
agg_op = Max | Min | Count | Sum | Avg
cmp_op = Equal | GreaterThan | LessThan | Other
```

[Yin 2017, Yin 2018, Rabinovich 2017]
Sentiment Analysis

- interpretation and classification of emotions (positive, negative and neutral) within text data using text analysis techniques
Sentiment Analysis

- XLNet based classification

*XLNet = GPT (AR)+BERT(AE): permutation AR (Transformer-XL)
RNN for NLP - softmax is often only calculated at the last step

CNN for NLP
ConvNet for NLP

- CNN architecture for sentence classification

Text classification by CNN+LSTM

- CNN: advantages in selecting good features
- LSTM networks: good abilities of learning sequential data.
Style-based fake news detection
Wording, writing style

NAVER cleanbot 2.0– Detecting malicious comments

1) Use transfer learning
   : improve the performance of the model by training other related tasks before training target task
   : Use ELMo for transfer learning

2) Use CNN+BiLSTM+LSTM
   for malicious comments detection

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>no transfer learning</td>
<td>0.9503</td>
<td>0.8949</td>
<td>0.9495</td>
<td>0.9214</td>
</tr>
<tr>
<td>ELMo</td>
<td>0.9669</td>
<td>0.9269</td>
<td>0.9686</td>
<td>0.9473</td>
</tr>
</tbody>
</table>

=> Better performance when using ELMO as transfer learning

https://d2.naver.com/helloworld/7753273
Cross-Lingual Adversarial Method

ROAD (Reference Oriented Adversarial Fake News Detector)
MRC-QA: SQuAD2.0

- Unanswerable question (negative example)
  - Relevant to the topic
  - Existence of plausible answers

<table>
<thead>
<tr>
<th>Article: Endangered Species Act</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paragraph: “... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”</td>
</tr>
</tbody>
</table>

| Question 1: “Which laws faced significant opposition?” |
| Plausible Answer: later laws |

| Question 2: “What was the name of the 1937 treaty?” |
| Plausible Answer: Bald Eagle Protection Act |

Figure 1: Two unanswerable questions written by crowdworkers, along with plausible (but incorrect) answers. Relevant keywords are shown in blue.
SQuAD2.0

- BERT – no answer prediction
MRC - HotpotQA

- multi-hop (two-hop) reasoning
- multiple supporting doc
- sentence-level supporting facts


**HGN**

![Diagram of HGN](image)
Reasoning MRC Dataset

✓ Quoref (Questions Requiring Coreferential Reasoning)
- Containing 24K questions over 4.7K paragraphs from Wikipedia.
- Resolve hard coreferences before selecting the appropriate span(s) in the paragraphs for answering questions.

✓ DROP (Discrete Reasoning Over Paragraphs)
- Wikipedia-based 96,000 QA pairs over 6,700 paragraphs.
- Most of the questions are mainly based on numerical arithmetic.

<Quoref>

<table>
<thead>
<tr>
<th>Reasoning</th>
<th>Passage (some parts shortened)</th>
<th>Question</th>
<th>Answer</th>
<th>BiDAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtraction (28.8%)</td>
<td>That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was sold by Robert Lehman for $16.3 million, well above its $12 million high estimate.</td>
<td>How many more dollars was the Untitled (1981) painting sold for than the 1.2 million dollar estimation?</td>
<td>$400000</td>
<td>$16.3 million</td>
</tr>
</tbody>
</table>

| Comparison (18.8%) | In 1517, the seventeen-year-old King sailed to Castle. There, the Flemish court ... In May 1518, Charles traveled to Barcelona in Aragon. | Where did Charles travel to first, Castle or Barcelona? | Castle | Aragon |

| Selection (19.4%) | In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack to tell the story of the events that led up to the battle. | Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller? | Don Mueller | Baker |

| Addition (11.7%) | Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalk, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on 2 March 1992. The JNA formed a battlegroup to counterattack the next day. | What date did the JNA form a battlegroup to counterattack after the village of Nos Kalk was captured? | 3 March 1992 | 2 March 1992 |

| Count (16.5%) and Sort (11.7%) | Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal. Carolina closed out the half with Kasay nailing a 44-yard field goal. ... In the fourth quarter, Carolina sealed the win with Kasay’s 42-yard field goal. | Which kicker kicked the most field goals? | John Kasay | Matt Prater |


| Other Arithmetic (3.2%) | Although the movement initially gathered some 60,000 adherents, the subsequent establishment of the Bulgarian Exarchate reduced their number by some 75%. | How many adherents were left after the establishment of the Bulgarian Exarchate? | 15000 | 60,000 |

| Set of spans (6.0%) | According to some sources 363 civilians were killed in Kavadarci, 230 in Negotino and 40 in Vatasha. | What does AFR stand for? | Kavadarci, Negotino, Vatasha, one of the Big Four audit firms |

| Other (6.8%) | This Annual Financial Report is our principal financial statement of accountability. The AFR gives a comprehensive view of the Department’s financial activities ... | What were the 3 villages that people were killed in? | Annual Financial Report | Vatasha | Negotino | Kavadarci | Vatasha | One of the Big Four audit firms |

<DROP>

<table>
<thead>
<tr>
<th>Reasoning Type</th>
<th>Example(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferring the entity to complete the 2nd-hop question (Type I)</td>
<td>Paragraph A: The 2015 Diamond Head Classic was a college basketball tournament ... Buddy Hield was named the tournament’s MVP. Paragraph B: Charano Rainier “Buddy” Hield is a Bahamian professional basketball player for the Sacramento Kings of the NBA... Q: Which team does the player named 2015 Diamond Head Classic’s MVP play for?</td>
</tr>
</tbody>
</table>

| Comparing two entities (Comparison) | Paragraph A: Lost Alone were a British rock band ... consisted of Steven Batelle, Alan Williamson, and Mark Gibson... Paragraph B: Guster is an American alternative rock band ... Founding members Adam Gardner, Ryan Miller, and Brian Rosemary... began... Q: Did Lost Alone and Guster have the same number of members? (yes) |

| Locating the answer by checking multiple properties (Type II) | Paragraph A: Several current and former members of the Pittsburgh Pirates ... John Milner, Dave Parker, and Rod Scurry... Paragraph B: David Gene Parker, nicknamed "The Cobra", is an American former player in Major League Baseball... Q: Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"? |

| Inferring about the property of an entity in question through a bridge entity (Type III) | Paragraph A: Marine Tactical Air Command Squadron 28 is a United States Marine Corps aviation command and control unit based at Marine Corps Air Station Cherry Point... Paragraph B: Marine Corps Air Station Cherry Point ... is a United States Marine Corps airfield located in Havelock, North Carolina, USA... Q: What city is the Marine Air Control Group 28 located in? |

| Other types of reasoning that require more than two supporting facts (Other) | Paragraph A: The towns of Yodobashi, Okubo, Totsuka, and Ochitai town were merged into Yodobashi ward. ... Yodobashi Camera is a store with its name taken from the town and ward. Paragraph B: Yodobashi Camera Co., Ltd. is a major Japanese retail chain specializing in electronics, PCs, cameras and photographic equipment. Q: Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese retail chain specializing in electronics, PCs, cameras, and photographic equipment it’s name? |
Framework

- **Multiple span type prediction**
  - Single span type prediction: Start and end probability for each token is computed.
  - Classification for count type prediction: 10 class classification problem (0~9), which covers about 97% counting problem in the DROP dataset.
  - Classification for add/sub type prediction: only addition and subtraction operations are involved, each number is classified into one of (-1, 0, +1).

- **Answer type prediction**
  - The answers to the questions are categorized into 5 types.

- **Question Directed Graph Attention Network**
  - Relation information between numbers and entities in the passages is updated through graph reasoning (self-attention mechanism).

- **RoBERTa Encoder**
  - Question and passage tokens are encoded by large RoBERTa encoder (max sequence: 512).
  - Mean pooling used to direct the reasoning in graph reasoning.

**Answer Type Prediction (FFN)**

- Single span type prediction:
  - Start and end probability for each token is computed.

- Multiple span type prediction:
  - Each token is classified by I,O tagging method, and multiple tokens classified as I are determined as answers.

- Classification for count type prediction:
  - 10 class classification problem (0~9), which covers about 97% counting problem in the DROP dataset.

- Classification for add/sub type prediction:
  - only addition and subtraction operations are involved, each number is classified into one of (-1, 0, +1)
Graph Reasoning Module

Node
✓ Entities (noun & pronoun) & numbers in sentences
   - Each number type is categorized into 7 number types.
     (CARDINAL, PERCENT, MONEY, TIME, DATE, ORDINAL, QUANTITY)

Edge = attention strength
✓ The numbers of the same type are connected with each other by the type-specific edge.
✓ The entities and the numbers are connected when they co-occur in a sentence.

Graph Attention Network

1. Numerical Reasoning Module
   → The passage and question representation from RoBERTa’s output are linearly transformed and input into the graph reasoning module.

2. Question Directed Node Embedding Update
   → Query, key, value are calculated for graph attention

3. Directed Graph Attention
   → After applying attention to the representations of neighboring nodes, the representation of the corresponding node is updated.

4. Module Output
   → Representations of entities and numbers are updated.

*Only number, entity updated

Question directed graph attention network (step 2,3 repeated)

Q: Question
P: passage

: Self-attention in graph
CoQA

- large-scale dataset which contains questions that depend on a conversation history
- Answers can be **free-form text**, not text spans from the passages (first select a text span as the rationale and then edit it to obtain a free-form answer)
- Seven domains passage

## 2. Dataset

<table>
<thead>
<tr>
<th>Domain</th>
<th>#Passages</th>
<th>#Q/A pairs</th>
<th>Passage length</th>
<th>#Turns per passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children’s Sto.</td>
<td>750</td>
<td>10.5k</td>
<td>211</td>
<td>14.0</td>
</tr>
<tr>
<td>Literature</td>
<td>1,815</td>
<td>25.5k</td>
<td>284</td>
<td>15.6</td>
</tr>
<tr>
<td>Mid/High Sch.</td>
<td>1,911</td>
<td>28.6k</td>
<td>306</td>
<td>15.0</td>
</tr>
<tr>
<td>News</td>
<td>1,902</td>
<td>28.7k</td>
<td>268</td>
<td>15.1</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1,821</td>
<td>28.0k</td>
<td>245</td>
<td>15.4</td>
</tr>
<tr>
<td>*only test data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reddit</td>
<td>100</td>
<td>1.7k</td>
<td>361</td>
<td>16.6</td>
</tr>
<tr>
<td>Science</td>
<td>100</td>
<td>1.5k</td>
<td>251</td>
<td>15.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>8,399</td>
<td>127k</td>
<td>271</td>
<td>15.2</td>
</tr>
</tbody>
</table>

Table 2: Distribution of domains in CoQA.

<table>
<thead>
<tr>
<th></th>
<th>SQuAD</th>
<th>CoQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Passage Length</strong></td>
<td>117</td>
<td>271</td>
</tr>
<tr>
<td><strong>Question Length</strong></td>
<td>10.1</td>
<td>5.5</td>
</tr>
<tr>
<td><strong>Answer Length</strong></td>
<td>3.2</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Table 3: Average number of words in passage, question and answer in SQuAD and CoQA.

<table>
<thead>
<tr>
<th></th>
<th>SQuAD</th>
<th>CoQA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Answerable</strong></td>
<td>66.7%</td>
<td>98.7%</td>
</tr>
<tr>
<td><strong>Unanswerable</strong></td>
<td>33.3%</td>
<td>1.3%</td>
</tr>
<tr>
<td><strong>Span found</strong></td>
<td>100.0%</td>
<td>66.8%</td>
</tr>
<tr>
<td><strong>No span found</strong></td>
<td>0.0%</td>
<td>33.2%</td>
</tr>
<tr>
<td><strong>Named Entity</strong></td>
<td>35.9%</td>
<td>28.7%</td>
</tr>
<tr>
<td><strong>Noun Phrase</strong></td>
<td>25.0%</td>
<td>19.6%</td>
</tr>
<tr>
<td><strong>Yes</strong></td>
<td>0.0%</td>
<td>11.1%</td>
</tr>
<tr>
<td><strong>No</strong></td>
<td>0.1%</td>
<td>8.7%</td>
</tr>
<tr>
<td><strong>Number</strong></td>
<td>16.5%</td>
<td>9.8%</td>
</tr>
<tr>
<td><strong>Date/Time</strong></td>
<td>7.1%</td>
<td>3.9%</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td>15.5%</td>
<td>18.1%</td>
</tr>
</tbody>
</table>

Table 4: Distribution of answer types in SQuAD and CoQA.

Figure 1: A conversation from the CoQA dataset. Each turn contains a question (Qt), an answer (At) and a rationale (Rt) that supports the answer.
1. Characteristics
- Coreference to previous questions and answers
- **Open-ended questions** (Yes, No, unknown answers)
- **Dialog action (teacher -> student)**
  (1) continuation (follow up, maybe follow up, don’t follow up)
  (2) affirmation (yes, no, neither)
  (3) answerable (answerable, no answer)
- **Answer = (answer span, dialog action)**

2. Dataset (**Metrics: F1, HEQ-Q, HEQ-D**)

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev.</th>
<th>Test</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>questions</td>
<td>83,568</td>
<td>7,354</td>
<td>7,353</td>
<td>98,407</td>
</tr>
<tr>
<td>dialogs</td>
<td>11,567</td>
<td>1,000</td>
<td>1,002</td>
<td>13,594</td>
</tr>
<tr>
<td>unique sections</td>
<td>6,843</td>
<td>1,000</td>
<td>1,002</td>
<td>8,854</td>
</tr>
<tr>
<td>tokens / section</td>
<td>396.8</td>
<td>440.0</td>
<td>445.8</td>
<td>401.0</td>
</tr>
<tr>
<td>tokens / question</td>
<td>6.5</td>
<td>6.5</td>
<td>6.5</td>
<td>6.5</td>
</tr>
<tr>
<td>tokens / answer</td>
<td>15.1</td>
<td>12.3</td>
<td>12.3</td>
<td>14.6</td>
</tr>
<tr>
<td>questions / dialog</td>
<td>7.2</td>
<td>7.4</td>
<td>7.3</td>
<td>7.2</td>
</tr>
<tr>
<td>% yes/no</td>
<td>26.4</td>
<td>22.1</td>
<td>23.4</td>
<td>25.8</td>
</tr>
<tr>
<td>% unanswerable</td>
<td>20.2</td>
<td>20.2</td>
<td>20.1</td>
<td>20.2</td>
</tr>
</tbody>
</table>

**The human equivalence score (HEQ)**: judge whether a system’s output is as good as that of an average human

- **HEQ-Q**: the percentage of questions for which this is true
- **HEQ-D**: the percentage of dialogs for which this is true for every question in the dialog

---

Figure 1: An example dialog about a Wikipedia section. The student, who does not see the section text, asks questions. The teacher provides a response in the form of a text span (or No answer), optionally yes or no (Yes / No), and encouragement about continuing a line of questioning (should, could, or should not ask a follow-up question).
Focus on handling **conversation history** to understand and answer the current question

- History attention mechanism (compute the attention weight according to the **utility of the history turn** in each variation)

- Positional history answer embedding (PosHAE) – multi-task learning (MTL)

- MTL (answer span prediction & dialog act prediction)

**Dataset**: QuAC  
**Metrics**: F1, HEQ-Q, HEQ-D

<table>
<thead>
<tr>
<th>Models</th>
<th>F1</th>
<th>HEQ-Q</th>
<th>HEQ-D</th>
<th>Yes/No</th>
<th>Follow up</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiDAF++</td>
<td>51.8/50.2</td>
<td>45.3/43.3</td>
<td>2.0/2.2</td>
<td>86.4/85.4</td>
<td>59.7/59.0</td>
</tr>
<tr>
<td>BiDAF++ w/ 2-C</td>
<td>60.0/60.1</td>
<td>55.7/54.8</td>
<td>5.3/4.0</td>
<td>86.6/85.7</td>
<td>61.0/61.3</td>
</tr>
<tr>
<td>BERT + HAE</td>
<td>63.9/62.4</td>
<td>59.7/57.8</td>
<td>5.9/5.1</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>FlowQA</td>
<td>64.6/64.1</td>
<td>-/59.6</td>
<td>-/5.8</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>BERT + PosHAE</td>
<td>64.7/ -</td>
<td>60.7/-</td>
<td>-/6.0/-</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>HAM</td>
<td>65.7/-/64.4</td>
<td>62.1/60.2</td>
<td>7.3/6.1</td>
<td>88.3/88.4</td>
<td>62.3/61.7</td>
</tr>
<tr>
<td>HAM (BERT-Large)</td>
<td>66.7/-/65.4</td>
<td>63.3/61.8</td>
<td>9.5/6.7</td>
<td>88.2/88.2</td>
<td>62.4/61.0</td>
</tr>
</tbody>
</table>

Each cell displays val/test scores.

variation: **only one turn of** conversation history (q, a pair sequence)

*Sliding window for long passage

*history

q: question, p: passage, H: history
### CoQA

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>In-domain</th>
<th>Out-of-domain</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance  Stanford University (R Orden &amp; Chen et al. TACL '19)</td>
<td>89.4</td>
<td>87.4</td>
<td>88.0</td>
</tr>
<tr>
<td>2</td>
<td>RoBERTa + AT + KD (ensemble)  Zhulu Technology</td>
<td>91.4</td>
<td>89.2</td>
<td>90.7</td>
</tr>
<tr>
<td>3</td>
<td>TR-MT (ensemble)  WeChatAI</td>
<td>91.5</td>
<td>88.8</td>
<td>90.7</td>
</tr>
<tr>
<td>4</td>
<td>RoBERTa + AT + KD (single model)  Zhulu Technology</td>
<td>90.9</td>
<td>89.2</td>
<td>90.4</td>
</tr>
<tr>
<td>5</td>
<td>TR-MT (ensemble)  WeChatAI</td>
<td>91.1</td>
<td>87.9</td>
<td>90.2</td>
</tr>
<tr>
<td>6</td>
<td>Google SQuAD 2.0 + MmFT (ensemble)  NSRAS + SDRG</td>
<td>89.9</td>
<td>88.0</td>
<td>89.4</td>
</tr>
<tr>
<td>7</td>
<td>TR-MT (single model)  WeChatAI</td>
<td>90.4</td>
<td>86.8</td>
<td>89.3</td>
</tr>
<tr>
<td>8</td>
<td>XLNet + Augmentation (single model)  XunMing</td>
<td>89.9</td>
<td>86.9</td>
<td>89.0</td>
</tr>
<tr>
<td>9</td>
<td>Google SQuAD 2.0 + MmFT (single model)  NSRAS + SDRG</td>
<td>88.5</td>
<td>86.0</td>
<td>87.8</td>
</tr>
<tr>
<td>10</td>
<td>ConVoR (ensemble)  Joint Laboratory of HIT and IFLYTEK Research</td>
<td>88.7</td>
<td>85.4</td>
<td>87.5</td>
</tr>
<tr>
<td>11</td>
<td>BERT + MmFT + ADA (ensemble)  Microsoft Research Asia</td>
<td>87.5</td>
<td>85.3</td>
<td>86.8</td>
</tr>
<tr>
<td>12</td>
<td>BERT + MmFT + ADA (single model)  Microsoft Research Asia</td>
<td>87.7</td>
<td>84.6</td>
<td>86.8</td>
</tr>
<tr>
<td>13</td>
<td>BERT (single model)  Microsoft Research Asia</td>
<td>86.4</td>
<td>81.9</td>
<td>85.0</td>
</tr>
<tr>
<td>14</td>
<td>HAM (single model)  University of Massachusetts, Amherst, and New Jersey, Rutgers University</td>
<td>85.7</td>
<td>81.7</td>
<td>84.6</td>
</tr>
</tbody>
</table>

### SQuAD2.0

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance  Stanford University (Rajpurkar &amp; Jia et al. 2018)</td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td>2</td>
<td>SA-Net on Albert (ensemble)  QIANXIN</td>
<td>90.724</td>
<td>93.011</td>
</tr>
<tr>
<td>3</td>
<td>SA-Net-V2 (ensemble)  QIANXIN</td>
<td>90.679</td>
<td>92.948</td>
</tr>
<tr>
<td>4</td>
<td>Retro-Reader (ensemble)  Shanghai AI Research</td>
<td>90.978</td>
<td>92.978</td>
</tr>
<tr>
<td>5</td>
<td>EntitySpanFocusV2 (ensemble)  RICOH_SRCB_DML</td>
<td>90.826</td>
<td>92.826</td>
</tr>
<tr>
<td>6</td>
<td>ATIRL+PV (ensemble)  Hithink RoyalFlush</td>
<td>90.442</td>
<td>92.877</td>
</tr>
<tr>
<td>7</td>
<td>ELECTRA+ALBERT+EntitySpanFocus (ensemble)  SRCB_DML</td>
<td>90.442</td>
<td>92.839</td>
</tr>
<tr>
<td>8</td>
<td>ELECTRA+ALBERT+EntitySpanFocus (ensemble)  SRCB_DML</td>
<td>90.429</td>
<td>92.799</td>
</tr>
<tr>
<td>9</td>
<td>EntitySpanFocus+AT (ensemble)  RICOH_SRCB_DML</td>
<td>90.454</td>
<td>92.748</td>
</tr>
<tr>
<td>10</td>
<td>ALBERT + DAFF + Verifier (ensemble)  PINGEN Omnis-Sentic</td>
<td>90.366</td>
<td>92.777</td>
</tr>
<tr>
<td>11</td>
<td>Retro-Reader on ALBERT (ensemble)  Shanghai AI Research</td>
<td>90.115</td>
<td>92.580</td>
</tr>
<tr>
<td>12</td>
<td>ELECTRA+EntitySpanFocus (ensemble)  Hithink RoyalFlush</td>
<td>90.002</td>
<td>92.497</td>
</tr>
<tr>
<td>13</td>
<td>ALBERT + DAFF + Verifier (ensemble)  PINGEN Omnis-Sentic</td>
<td>90.002</td>
<td>92.425</td>
</tr>
<tr>
<td>14</td>
<td>ALBERT (ensemble)  Google Research &amp; IITC</td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td>15</td>
<td>ALBERT (ensemble)  Google Research &amp; IITC</td>
<td>89.743</td>
<td>92.180</td>
</tr>
<tr>
<td>16</td>
<td>ELECTRA+ATIRL+PV (single model)  Hithink RoyalFlush</td>
<td>89.551</td>
<td>92.366</td>
</tr>
</tbody>
</table>
The Stanford Natural Language Inference (SNLI) corpus -
https://nlp.stanford.edu/projects/snli/

- Size: 570k sentence pairs
- manually labeled for balanced classification with the labels *entailment*, *contradiction*, and *neutral*, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE)

- **SoTA model**: CA-MTL
- **Metric**: Accuracy

Trend on SNLI corpus
The Multi-Genre Natural Language Inference (MultiNLI) corpus - https://cims.nyu.edu/~sbowman/multinli/

- **Size**: 433k sentence pairs
- **crowd-sourced collection annotated with textual entailment information**
- **SoTA model**: T5-11B
- **Metric**: Accuracy

Transfer Learning with a Unified Text-to-Text Transformer

Trend on MNLI corpus
**Architecture**

*Conditionally adaptive multi-task learning

**CA-MTL (2020)**

26 tasks: 9 GLUE tasks, 8 Super-GLUE Tasks, 6 MRQA tasks, etc

*9 GLUE task
MNLI: multi-genre natural language inference (entail, contradict, neutral)
QQP: quora question pairs (semantic equivalence)
QNLI: question answering NLI (context sentence contains answer to the question)
SS-2: stanford sentiment treebank
cola: corpus of linguistic acceptability (grammaticality check)
sts-b: semantic textual similarity benchmark
mrpc: microsoft research paraphrase corpus (semantic equivalence)
rte: recognizing textual entailment (entail, not-entail)
WNLI: The Winograd Schema Challenge (common sense pronoun resolution task)

**Figure 1:** CA-MTL$_{\text{BASE}}$ architecture first uses our uncertainty-based sampling algorithm to choose task data for batching. Then, the input tokens go through a frozen embedding layer, followed by a Conditional Alignment layer. The rest contains frozen BERT-based Transformer layers and trainable adapters.
Grammatical Error Correction (GEC)

- Detecting and correcting grammatical errors from the sentence
  - Spelling errors
  - Wrong words
  - Bad grammars

**Writing assistant**

**Dialogue assistant**

---

**Common Mistakes in English**

- Lizzie approached me, say, "Hello!"
- Lizzie approached me and said, "Hello!"

- If someday we meet, would we again start?
- If someday we meet, would we start again?

- He neither has talent nor the desire to learn.
- He has neither talent nor the desire to learn.
Grammatical Error Correction (GEC)

- Neural machine translation approaches
  - Transformer and copy mechanism
  - 0.5880 $F_{0.5}$ score

Figure 1: Copy-Augmented Architecture.

Zhao et al., NAACL 2019
GEC Data Augmentation

• Generate wrong sentences from correct sentences
• Generated (wrong, correct) pairs are augmented to train the model
• The proposed methods can **generate without** annotated data
Model-based Data Generation

- Data generation model is trained with unlabeled data pair \((\tilde{y}, \tilde{y})\)
- The model generates subtly wrong sentences
- Appropriate generated pairs are used to train the GEC model
Multi-domain Task-oriented Dialogue

• The system interacts with user to help the user achieve his/her goal
  • e.g. Restaurant reservation, hotel reservation...

• Specific domains & specific goals
  • ⇔ Open domain (chit-chat)

• Multi-turns

• User goals are not limited to just one domain

To book a hotel

To book a restaurant near the hotel

To book a taxi from the hotel to the restaurant
The most important subtask of task-oriented dialogue system
Dialogue state generally consists of slot-value pairs
  - Slot means general class like food, area / value means specific value of slot
In multi-domain, dialogue state consists of domain-slot-value triplets
  - ex) (restaurant-pricerange-cheap), (attraction-type-museum)...
Infer dialogue state each turn of dialogues between user and system
  - There are ground truth dialogue states every turn
Inferred state is used to generate next system action and response

TRADE (TRAnsferable Dialogue statE generator)
  - 2019 ACL @ Hong Kong Univ. & Salesforce
    [https://www.aclweb.org/anthology/P19-1078]

DS-DST (Dual Strategy for DST)
  - State-Of-The-Art in DST
Dataset: MultiWOZ

- Standard benchmark dataset for end-to-end task-oriented dialogue systems
  - Multi-domain dataset for task-oriented dialog system
  - Virtual conversations between a tourist and a guide in Cambridge
  - 7 domains: Hotel, restaurant, attraction, train, taxi, police, hospital
  - Number of dialog: 10,000 / Avg. number of turns per dialog: 13

- Evaluation in MultiWOZ
  - Inform rate
    - How many entries provided by the system meet goals?
  - Success rate
    - How many user requests are provided by the system?
  - BLEU
    - Similarity between generated response and true response in the dataset
A short example in MultiWOZ

**User:** I am planning a trip to town and want to sight see a bit.
**System:** There are 79 attractions in the city. Do you have a specific type of attraction?

**User:** I’d like a architecture.
**System:** How about [All Saints Church]? It has fantastic architecture and free entrance.

**User:** Could I get the address? I would also like an expensive place to eat around it.
**System:** The address is Jesus Lane. What type of food would you like to eat?

**User:** I'm in the mood for Turkish food.
**System:** Anatolia is an expensive Turkish restaurant in the city centre.

**User:** Can I have the phone number?
**System:** The phone number is 01223362372.

**User:** Thank you!
**System:** You're welcome, feel free to contact us about anything else you need.
traditional pipeline architecture

User

“I want to find a Chinese restaurant.”

Natural Language Understanding

Inform (cuisine=“Chinese”)

“Where do you want to eat?”

Natural Language Generation

Dialog State

Request (location)

Dialog Policy

Dialog Manager

Knowledge Base

Query
Existing model: SOLOIST

- An auto-regressive model for training (Language modeling)
- Task 1: Predict dialogue state (slot-value)
- Task 2: Predict system response
- Task 3 (for Auxiliary loss)
  - Replace dialogue state or system response in input sequence into negative samples randomly
  - Then, predict input sequence is negative sample or not (binary classification)
- Jointly train by sum of 3 losses
DAMD (Domain Aware Multi-Decoder)

User utterance → Delexicalization → GRU encoder → Attention → User context → Attention → User context → Concat → GRU decoder → Softmax → New belief → DB

System Response → GRU encoder → Attention → Action context → Attention → Response context → Concat → GRU decoder → Softmax → New action → DB

Belief → GRU encoder → Attention → Action context → Attention → Belief context → Concat → GRU decoder → Softmax → New response → DB
Domain state tracking

• For tracking the flow of conversation from domain perspective
• Contains binary values indicating whether domains are activated or not in the current conversation
• Changes over turns

User: I would like to get some information about a restaurant and a hotel.
System: Okay, let's start with a hotel, any preference of type, area, or price range?

User: The hotel that I am looking for is called Gonville.
System: Do you want to book the hotel?

User: I want an Italian restaurant near the hotel.
Architecture w/ domain state tracking

- NLU module is shared for DST, POL, and NLG
- Darker blocks mean previous turn
- DB result contains the number of matched entries for each domain

*for POL GRU training use two level: use CE loss for Supervised Learning (SL); use reward (success rate and correct system action rate) for 2nd level Reinforcement Learning (RL) training

Cross entropy loss

Binary cross entropy loss
End-to-end ASR

Audio Inputs

Frontend (Preprocessing) STFT, MEL

Short-Time Fourier Transform (STFT)

Transformer Encoder

Transformer Decoder (Attention)

Speech Representation

Beam Search

CTC

Text Outputs

*Connectionist Temporal Classification (CTC)
**Conformer Encoder**

**Conformer (Encoder)**

**Decoder - LSTM**


*convolution-augmented transformer for speech recognition, Conformer

*SpecAugment: A simple data augmentation method for automatic speech recognition (to log-mel spectrogram): time warping, frequency & time masking
Conformer + Wav2Vec 2.0 (Encoder)  
Decoder - LSTM  

*Pre-training: The contrastive loss between the context vectors from masked features and the quantization unit is optimized.

CTC-attention decoder

*use intermediate label representation
allowing repetitions of labels and blank labels
Do baum-welch optimization for Viterbi alignment

Joint CTC-Attention Decoding

LAS (Listen, Attend and Spell)

Grapheme characters \( y_i \) are modelled by the Character Distribution

*Attention-based recurrent network decoder (character decoding)

*LAS does not make independence assumptions in the label sequence unlike CTC

*accepts filter bank spectra as inputs

Tacotron2: Seq2seq with attention RNN + modified WaveNet

*Wavenet: invert the mel spectrogram feature representation into time-domain
  waveform samples (10 component mixture of logistic distributions (MoL) to generate 16-bit
  samples at 24 kHz)

*Location-sensitive attention: mitigating potential failure modes where some subsequences are repeated or ignored by the decoder
*Auto-regressive decoder to generate mel spectrogram frame by frame; Previous frame input to pre-net & post-net (for residual)
*Stop token to stop generation
MelGAN

Change WaveNet vocoder to:
- MelGAN for faster inference

Figure 1: MelGAN model architecture. Each upsampling layer is a transposed convolution with kernel-size being twice of the stride (which is same as the upsampling ratio for the layer). 256x upsampling is done in 4 stages of 8x, 8x, 2x and 2x upsampling. Each residual dilated convolution stack has three layers with dilation 1, 3 and 9 with kernel-size 3, having a total receptive field of 27 timesteps. We use leaky-relu for activation. Each discriminator block has 4 strided convolution with stride 4. Further details can be found in the Appendix 6.
Dialog System for Counseling (TOD → Open-ended)

- Paraphrasing
- Reflect Feeling
- Correct Information
- Logical Answering
- Questioning
- Show Interest
- Listening
- Memorizing
CBT Counseling Scenario

Initiation
- Greeting
- Explanation of Counseling Method
- Extraction information of User

Situation
- 5W1H details
- Automatic thoughts
- Ground Belief
  - Emotion
  - Supporting Evidence
  - No-Supporting Evidence

Refutation
- Alternative thoughts
- Emotion
- Motivating

*Automatic vs alternative

*For next visit
VHRED: RNN Dialog Model

*Variational hierarchical recurrent

User: How are things?
Meena: They’re good. How about you?
User: I’m trying to decide what to watch
Meena: What are some of your favorite shows?
User: I like the Good Place
Meena: I haven’t watched that. My favorite shows are Supernatural, and Star trek: TNG.
User: What is TNG?
Meena: The Next Generation
Frederick Parker-Rhodes (21 March 1914 - 21 November 1987) was an English linguist, plant pathologist, computer scientist, mathematician, mystic, and mycologist.

Data2Text NLG requires **accuracy (semantic fidelity)** as well as fluency.
Switch-GPT2 (Chen et al., 2020)

\[ L = L_c + \lambda \sum_{w_j \in m} (1 - P_{copy}^j) \]

*\( L_c \) original loss \( w \), \( V_i \) slot-value, \( w_j \) target word in slot-value for copy.
• KGPT (Chen et al., 2020)

The encoder is pre-trained transformer model (specially trained for data2text task)

\[ P(y_i = w) = p_{gen} P_{voc}(w) + (1 - p_{gen}) \sum_{j: x_j = w} \alpha_j \]

*Fine tuned w/ any table2text

*Knowledge

*Transformer encoder/decoder
Both pre-trained D2T Architecture

ToTTo dataset (Parikh et al., 2020): Wikipedia tables paired with natural language descriptions 120K

\[
\text{loss}_t = -\log P(w_t^*) + \lambda \sum_i \min(a_t^i, c_t^i)
\]

Duplicate Copy penalty

\[
c^t = \sum_{t'=0}^{t-1} d^t
\]
- NLP enjoys rapid progress in the last 10 years due to deep learning.

- Even more rapid progress in the last few years due to larger models, better usage of unlabeled data

- NLP is reaching the point of having big social impact, making issues like bias and security increasingly important.

- Big model, big computation resources, huge training times are problematic, need to focus more light way of doing NLP (even embedded model)