Outline

Morning program

- Preliminaries
- Semantic matching
- Learning to rank
- Entities

Afternoon program

- Modeling user behavior
- Generating responses
- Recommender systems
- Industry insights
- Q & A

Deep Learning in industry

- Companies have endless amounts of data!
  Or do they?
- Performance
  Is .9 accuracy/F_1/etc. good enough?
  No? Would 0.95 be?
- Business logic/constraints
  - *Your model is doing great in general, but not in case X, Y and Z.*
  Can you keep it exactly as it is now, and fix just these cases?
- Explicit domain knowledge
  E.g.: recommending product X for user Y is not applicable, as it is not available where user Y lives.

Deep Learning in industry

- Hybrid Code Networks
  Combining RNNs with domain-specific knowledge
- Smart Reply
  Automated response suggestion for email

Hybrid Code Networks

Task

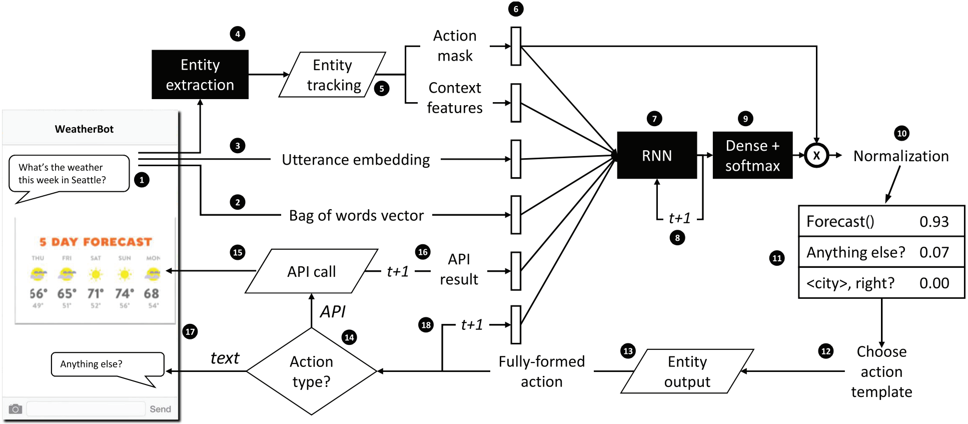
Dialogue system. User can converse with a system that can interact with APIs.

Combining RNNs with domain-specific knowledge

- Incorporate business logic by including modules in the system that can be programmed
- Explicitly condition actions on external knowledge

[Williams et al., 2017]

Hybrid Code Networks



What's the weather this week in Seattle?

Anything else?

Forecast()	0.93
Anything else?	0.07
<city>, right?	0.00

Trapezoids refer to programmatic code provided by the software developer.
Shaded boxes are trainable components.

[Williams et al., 2017]

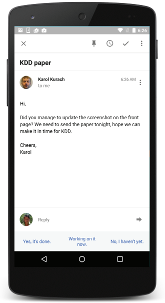
Smart Reply

Automated response suggestion for email

Use an RNN to generate responses for any given input message.

Additional constraints

- **Response quality**
  Ensure that the individual response options are always high quality in language and content.
- **Utility**
  Select multiple options to show a user so as to maximize the likelihood that one is chosen.
- **Scalability**
  Process millions of messages per day while remaining within the latency requirements.
- **Privacy**
  Develop this system without ever inspecting the data except aggregate statistics.



KDD paper

Karol Kurach
to me
6:26 AM

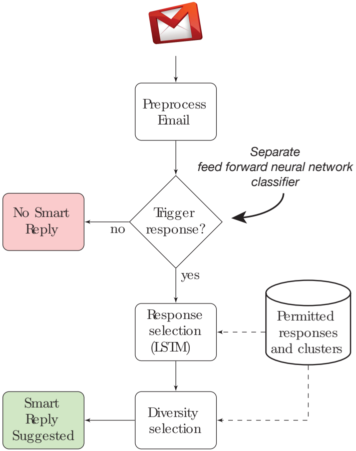
Hi,

Did you manage to update the screenshot on the front page? We need to send the paper tonight, hope we can make it in time for KDD.

Cheers,
Karol

[Kannan et al., 2016]

Smart Reply


Preprocess Email
Separate feed forward neural network classifier
Trigger response?
no
No Smart Reply
yes
Response selection (LSTM)
Permitted responses and clusters
Diversity selection
Smart Reply Suggested

[Kannan et al., 2016]

Response selection

- Construct a set of allowed responses R.
- Organise the elements of R into a trie.
- Conduct a left-to-right beam search, and only retain hypotheses that appear in the trie.

Complexity: O(beam size × response length).

Utility/diversity

Goal: present user with diverse responses

Instead of “No”, “No, thanks”, and “Thanks!”, we’d rather produce “No, thanks”, “Yes, please”, “Let me come back to it”.

- Manually label a couple of messages per response intent.
- Use a state-of-the-art label propagation algorithm to label all other messages in R.

What do we learn?

- Deep learning component is a (small) part of a much larger system.
- Getting the right training data can be hard.
- The machine learned part is guided/corrected/prevented from predicting undesired output.

Neural IR at Bing

Long history of neural IR models at Bing/Microsoft

- RankNet/LambdaRank [Burges et al., 2005, 2006]
- ListNet/ListMLE [Cao et al., 2007, Xia et al., 2008]
- DSSM/CDSSM [Huang et al., 2013, Shen et al., 2014]
- Recent representation learning models for long text [Mitra et al., 2017, Zamani et al., 2018]

NN and GBDT are both popularly used across many teams

Neural IR at Bing

Beyond Web search, heavy use of deep learning systems for

- Speech recognition [Xiong et al., 2017c]
- Conversational models (e.g., Cortana & Zo)
- Machine translation [Hassan et al., 2018]
- Machine reading [Wang et al., 2017] and emerging Office Intelligence scenarios (e.g., [Van Gysel et al., 2017])
- And others...

Neural IR at Bing

Some of the unique challenges and considerations:

- Supervision
  - Large (explicitly/implicitly) labeled datasets are available for training deep models in Web search
  - Not available for many multi-tenant enterprise scenarios due to privacy and scalability considerations—distance supervision and other approaches may be necessary
- Infrastructure investments
  - GPU and other machine resources for experimentation; serving infrastructure investments for running deep models in production
  - Neural model based features vs. rethinking the stack with neural models as first class citizens
- Model reuse: across tenants and different services