Deep Learning in Domain Scaling for Conversational Agents

Ye-Yi Wang

In collaboration with Bin Cao, Jianfeng Gao, Ruhi Sarikaya, Gokhan Tur, Asli Celikyilmaz, Bing Intent Science Team and MSR Deep Learning Center
Growing with interACT

Thanks for leading the community to shape the reality
Looking forward to continued leadership in shaping the future
Cortana: Task Completion & QA

- Bluetooth is now off.
- You could probably go without one. Here's the weather.
- SeaTac, Washington
  - 73°, WED 7/6 Low 59°, Mostly Sunny 10%
  - 60°, 9:00 AM, Mostly Sunny 0%
  - 66°, 12:00 PM, Mostly Cloudy 10%
  - 71°, 3:00 PM, Mostly Sunny 10%
  - 72°, 6:00 PM, Mostly Sunny 0%
  - 60°, 9:00 PM, Mostly Sunny 0%

- Do I need a jacket today?
- Try Is it raining in Miami?
- How to say "Hallo" in German

- Turn off Bluetooth please
- Web
  - Bing
  - English
    - Hello
  - German
    - Hallo

Data from Microsoft Translator
Cortana: Multi-turn Conversations

Remind me to buy milk

10 Pm tomorrow
Cortana: Language Understanding

• What is “Understanding”?
  • Explicit or implicit? Generic or domain specific
  • Practical solution: Query → Semantic Frame

• Semantic Frame: structured meaning representation
  • Domain (Weather, Device Control, Play Music, …) SVMs
  • Intent (5 day forecast, Get temperature, …) SVMs
  • Slots (e.g., weather in <loc>Boston</loc>) CRFs

• Model Training
  • Domain by domain, locale by locale
  • Annotators provide labeled data for initial coldstart model training
  • Annotators label the feedback data after deployment for continuous improvement
  • Hard to scale
Cortana: Dialog Modeling

• 1\textsuperscript{st} generation (past): manually designed finite state dialog flow/policy
• 2\textsuperscript{nd} generation (now): a platform that hides the complexity of flow design, fixed dialog policy
• 3\textsuperscript{rd} generation (future): deep reinforcement learning for dialog policy learning/tuning.
Why Language Understanding is hard

• Ambiguity
• Power Law

“there is no data like more data”
“data is the new oil, intelligence is the new power”
The Language Understanding Scaling Problem

• Domain scaling: a demand/supply problem of supervision data

• Increase the supply: Automatic offline data labeling & feedback loop
  • Multi-task deep learning for domain classification against an existing taxonomy (ODP)
  • HITS and EM algorithm for entity tagging
  • Feedback loop

• Reduce the demand
  • Features with better generalization capability (Multi-task embedding learning)
  • Models that generalize better (LSTM, Seq2Seq)
Increase the Supply

Tools for users to select from pre-labeling big data via semi-supervised or unsupervised learning
Semi-supervised/Unsupervised Labeling of Big Data

- Classification with weak supervisions

- Slot tagging with EM algorithms + Knowledge base
  - Substring match against entities in the knowledge base
  - Disambiguation via pattern statistics (contextual dependency)
  - Iteratively repeated the process (EM algorithm)
  - Initialize EM with HITS algorithm
Tools for data selection

• Browsable – organized according to a given domain taxonomy (ODP) and (finer-grained) clusters from topic modeling

• Searchable – semantic similarity ranking based on query embedding
Multi-Task Deep Learning
Multi-Task Deep Learning

DSSM
- Translation
  - Hidden3
  - Hidden2
  - Hidden1
  - Input
  - Hashing & Convolution
  - Ja-Jp Query

DSSM
- Click Pred
  - Hidden3
  - Hidden2
  - Hidden1
  - Input
  - Hashing & Convolution
  - Ads Listing

DNN
- Cls 1
  - Hidden3
  - Hidden2
  - Hidden1
  - Hashing & Convolution
  - En-US Query
- Cls 2
- Cls n

DSSM
- Click Pred
  - Hidden3
  - Hidden2
  - Hidden1
  - Hashing & Convolution
  - Web Doc
Multi-Task Deep Learning: learn generic semantics

• DNN/DSSM based multi-task learning has been applied to domain classification in IntentExplorer
• Significant improvement on Ads team’s ODP experiments

<table>
<thead>
<tr>
<th></th>
<th>Avg AUC</th>
<th>Top1 Accuracy</th>
<th>Top5 Accuracy</th>
<th>Top10 Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT-DNN</td>
<td>95.0%</td>
<td>51.3%</td>
<td>82.4%</td>
<td>89.5%</td>
</tr>
<tr>
<td>SVMs</td>
<td>95.6%</td>
<td>41.1%</td>
<td>72.3%</td>
<td>83.6%</td>
</tr>
</tbody>
</table>

• However, query level embedding doesn’t help slot tagging
Reduce the Demand

Embedding as features for better generalization
Embedding learning for Cold Start LU
Reducing the Demand on Labeled Data

On par performance can be achieved with a fraction of training data for slot tagging

<table>
<thead>
<tr>
<th>Domain</th>
<th>Baseline (Production Model)</th>
<th>Baseline (SVM + Ngram)</th>
<th>LSTM + Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>alarm</td>
<td>0.999</td>
<td>0.997</td>
<td>0.9995</td>
</tr>
<tr>
<td>calendar</td>
<td>0.997</td>
<td>0.992</td>
<td>0.9976</td>
</tr>
<tr>
<td>communication</td>
<td>0.996</td>
<td>0.976</td>
<td>0.9958</td>
</tr>
<tr>
<td>mediacontrol</td>
<td>0.999</td>
<td></td>
<td>0.9989</td>
</tr>
<tr>
<td>mystuff</td>
<td>0.997</td>
<td>0.997</td>
<td>0.9973</td>
</tr>
<tr>
<td>note</td>
<td>0.999</td>
<td>0.999</td>
<td>0.9995</td>
</tr>
<tr>
<td>ondevice</td>
<td>0.993</td>
<td></td>
<td>0.9944</td>
</tr>
<tr>
<td>places</td>
<td>0.989</td>
<td>0.984</td>
<td>0.9885</td>
</tr>
<tr>
<td>reminder</td>
<td>0.999</td>
<td>0.979</td>
<td>0.999</td>
</tr>
<tr>
<td>weather</td>
<td>0.999</td>
<td>0.998</td>
<td>0.9989</td>
</tr>
<tr>
<td>web</td>
<td>0.969</td>
<td>0.941</td>
<td><strong>0.9734</strong></td>
</tr>
<tr>
<td>webnavigation</td>
<td></td>
<td>0.998</td>
<td>0.9967</td>
</tr>
</tbody>
</table>

On par performance can be achieved without engineered features for domain classification
Opportunity for Improvement

Using Oracle embedding, the classification results were much better when fraction of training data were used.

<table>
<thead>
<tr>
<th>#Samples</th>
<th>Embedding</th>
<th>Optimal Embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>494</td>
<td>0.6197</td>
<td>0.8800</td>
</tr>
<tr>
<td>1080</td>
<td>0.7581</td>
<td>0.8972</td>
</tr>
<tr>
<td>2312</td>
<td>0.8418</td>
<td>0.9139</td>
</tr>
<tr>
<td>4974</td>
<td>0.8765</td>
<td>0.9290</td>
</tr>
</tbody>
</table>
Multi-Task Deep Sequence Learning

Query

Encoder

Classifier

ODP domain

Query with shared click

Decoder

word word word ... word

Clicked document title

Decoder

word word word ... word

Glossary in Dictionary

Decoder

word word word ... word

Cortana Slot Tagging

Decoder

word word slot ... word

Classifier

Language
Summary

• Challenges in scaling up or democratizing the conversational experiences

• The key issue is here is a demand/supply problem

• Increasing demand – auto-labeled data for selection

• Reducing the cost – project into a continuous space via embedding learning for better generalization