Cross-Domain Semantic Parsing via Paraphrasing

Yu Su & Xifeng Yan, EMNLP 2017
presented by Sha Li
Semantic Parsing

Mapping **natural language utterances** to **logical forms** that machines can act upon.

Example:

- **Database query**

  ```
  SELECT country.name
  FROM country, co2_emissions
  WHERE country.id = co2_emissions.country_id
  AND co2_emissions.year = 2014
  ORDER BY co2_emissions.volume DESC
  LIMIT 1;
  ```

- **Intents and arguments for a personal assistant**

  - **Angelina Jolie net worth**
    - (FactoidQuery
      - (Entity /m/0f4vbz)
      - (Attribute /person/net_worth))

  - **Play Sunny by Boney M**
    - (PlayMedia
      - (MediaType MUSIC)
      - (SongTitle "sunny")
      - (MusicArtist /m/017mh))

Database query

Intents and arguments for a personal assistant
In-domain VS Cross-domain Semantic Parsing

- **In-domain**: training/test set from the same domain
- **Cross-domain**: train on **source domain** and test on **target domain**
- **Why cross-domain**:
  - Sometimes we have more training data from one domain than another; collecting training data from the target domain is expensive
  - The source domain shares some similarities with the target domain, making it possible to train a cross-domain model
Challenges

1. Different domains have different logical forms (different predicate names etc.) ⇒ translate to a common middle ground: **canonical utterance**
   
   **Canonical utterance:** has a one-to-one mapping to the logical form

2. Vocabulary gap between domains ⇒ pretrained **word embeddings**

45%-70% of the words are covered by any of the other domains
Previous Work

Paraphrase based semantic parsing

Map utterances into a canonical natural language form before transforming into logical form. (Berant and Liang 2014, Wang et al. 2015)
The logical form is not shared across domains
Paraphrasing Framework

The paraphrase module is shared

The logical form is not shared across domains

In which seasons did Kobe Bryant play for the Lakers?

Season of player Kobe Bryant whose team is Los Angeles Lakers

R[season](player.KobeBryant \( \cap \) team.Lakers)

When did Alice start working for Mckinsey?

Start date of employee Alice whose employer is Mckinsey

R[start](employee.Alice \( \cap \) employer.Mckinsey)

External Language Resources
pre-trained word embeddings, monolingual parallel corpora, ...

Paraphrase Model

Input Utterance

Canonical Utterance

Logical Form
Problem Setting

- Assume that the mapping from canonical utterance to logical form is given for both domains
- Propose a seq2seq model for **paraphrasing**
- Use **pre-trained word embeddings** to help domain adaptation
  - Introduce standardization techniques to improve word embeddings
- **Domain adaptation** is done by: training a paraphrase model in the source domain and fine-tuning it the target domain
Paraphrase Model

Encoder-decoder structure.

The input of the decoder RNN at is the hidden state of the previous time step and the previous output.
Encoder-decoder with Attention

Attention vector:
weighted sum of the output from the encoder.

The input of the decoder RNN at is the hidden state of the previous time step, the previous output and the attention vector.
Analysis of Word Embeddings

300 dimension word2vec embeddings trained on the 100B word Google news corpus.

Compared to random initialization with unit variance:

- **Small micro variance**: the variance between dimensions of the same word is small

<table>
<thead>
<tr>
<th>Initialization</th>
<th>L2 norm</th>
<th>Micro Variance</th>
<th>Cosine Sim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>17.3 ± 0.45</td>
<td>1.00 ± 0.05</td>
<td>0.00 ± 0.06</td>
</tr>
<tr>
<td>WORD2VEC</td>
<td>2.04 ± 1.08</td>
<td>0.02 ± 0.02</td>
<td>0.13 ± 0.11</td>
</tr>
</tbody>
</table>
Analysis of Word Embeddings

300 dimension word2vec embeddings trained on the 100B word Google news corpus.

Compared to random initialization with unit variance:

- Small micro variance: the variance between dimensions of the same word is small
- Large macro variance: the L2 norm of different words varies largely

<table>
<thead>
<tr>
<th>Initialization</th>
<th>L2 norm</th>
<th>Micro Variance</th>
<th>Cosine Sim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>17.3 ± 0.45</td>
<td>1.00 ± 0.05</td>
<td>0.00 ± 0.06</td>
</tr>
<tr>
<td>WORD2VEC</td>
<td>2.04 ± 1.08</td>
<td>0.02 ± 0.02</td>
<td>0.13 ± 0.11</td>
</tr>
</tbody>
</table>
Embedding Standardization

- Per-example standardization: make variance of each row 1
  - Reduces variance of L2 norm among words
  - Cosine similarity between words is preserved
- Per-feature standardization: make the variance of each column 1
- Per-example normalization: make the L2 norm of each word 1

<table>
<thead>
<tr>
<th>Initialization</th>
<th>L2 norm</th>
<th>Micro Variance</th>
<th>Cosine Sim.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>17.3 ± 0.45</td>
<td>1.00 ± 0.05</td>
<td>0.00 ± 0.06</td>
</tr>
<tr>
<td>WORD2VEC</td>
<td>2.04 ± 1.08</td>
<td>0.02 ± 0.02</td>
<td>0.13 ± 0.11</td>
</tr>
<tr>
<td>WORD2VEC + ES</td>
<td>17.3 ± 0.05</td>
<td>1.00 ± 0.00</td>
<td>0.13 ± 0.11</td>
</tr>
<tr>
<td>WORD2VEC + FS</td>
<td>16.0 ± 8.47</td>
<td>1.09 ± 1.31</td>
<td>0.12 ± 0.10</td>
</tr>
<tr>
<td>WORD2VEC + EN</td>
<td>1.00 ± 0.00</td>
<td>0.01 ± 0.00</td>
<td>0.13 ± 0.11</td>
</tr>
</tbody>
</table>
Experiments: Dataset

Dataset contains 8 different domains.

The mapping from canonical utterances to logical forms are given.

The input utterances are collected via crowdsourcing.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Calendar</th>
<th>Blocks</th>
<th>Housing</th>
<th>Restaurants</th>
<th>Publications</th>
<th>Recipes</th>
<th>Social</th>
<th>Basketball</th>
</tr>
</thead>
<tbody>
<tr>
<td># of example ($N$)</td>
<td>837</td>
<td>1995</td>
<td>941</td>
<td>1657</td>
<td>801</td>
<td>1080</td>
<td>4419</td>
<td>1952</td>
</tr>
<tr>
<td># of logical form ($</td>
<td>\mathcal{Z}</td>
<td>$, $</td>
<td>\mathcal{C}</td>
<td>$)</td>
<td>196</td>
<td>469</td>
<td>231</td>
<td>339</td>
</tr>
<tr>
<td>vocab. size ($</td>
<td>\mathcal{V}</td>
<td>$)</td>
<td>228</td>
<td>227</td>
<td>318</td>
<td>342</td>
<td>203</td>
<td>256</td>
</tr>
<tr>
<td>% $\in$ other domains</td>
<td>71.1</td>
<td>61.7</td>
<td>60.7</td>
<td>55.8</td>
<td>65.6</td>
<td>71.9</td>
<td>46.0</td>
<td>45.6</td>
</tr>
<tr>
<td>% $\in$ WORD2VEC</td>
<td>91.2</td>
<td>91.6</td>
<td>88.4</td>
<td>88.6</td>
<td>91.1</td>
<td>93.8</td>
<td>86.9</td>
<td>86.9</td>
</tr>
<tr>
<td>% $\in$ other domains + WORD2VEC</td>
<td><strong>93.9</strong></td>
<td><strong>93.8</strong></td>
<td><strong>90.9</strong></td>
<td><strong>90.4</strong></td>
<td><strong>95.6</strong></td>
<td><strong>97.3</strong></td>
<td><strong>89.3</strong></td>
<td><strong>89.4</strong></td>
</tr>
</tbody>
</table>
Baselines

1. (Wang et al) Log-linear model.
2. (Xiao et al) Multi-layer perceptron to encode the unigrams and the bigrams of the input, and then use a RNN to predict the logical form.
3. (Jia and Liang) Seq2Seq model (bi-RNN with attentive decoder) to predict the linearized logical form.
4. (Herzig and Berant) Use all domains to train a single parser with a special encoding to differentiate between domains.
Random +I is the most basic model using random initialization of word embeddings.

This model is comparable to previous single domain models.
Experiments: Cross-Domain

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Herzig and Berant</td>
<td>79.6</td>
</tr>
<tr>
<td>Random</td>
<td>76.9</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>74.9</td>
</tr>
<tr>
<td>Word2Vec +EN</td>
<td>71.2</td>
</tr>
<tr>
<td>Word2Vec +FS</td>
<td>78.9</td>
</tr>
<tr>
<td>Word2Vec +ES</td>
<td>80.6</td>
</tr>
</tbody>
</table>

1. Directly using Word2Vec pretrained vectors hurts!
2. Per-example normalization (EN) decreases performance even more.
3. Both per-feature standardization (FS) and per-example standardization (ES) improves performance. Per-example standardization works better.

The performance gain is mainly due to word embedding standardization.
Other results

The improvement of cross-domain training is more significant when the target domain data is scarce.

The in-domain training data is downsampled.
Discussion on Standardization/Normalization

> Normalization improves performance in similarity tasks. (Levy et al. 2015)

> A word that is consistently used in a similar context will be represented by a longer vector than a word of the same frequency that is used in different contexts. The L2 norm is a measure of word significance. (Wilson and Schakel 2015)

It is worth trying different normalization schemes for your task!
Conclusion

1. The semantic parsing problem can be decomposed into two steps: first paraphrase the utterance into a canonical form, then translate this canonical form into logical form (idea from Berant and Liang, 2014)
2. Paraphrasing can be learned by a seq2seq model. (We can formulate paraphrasing as translation)
3. Initialization of word embeddings is critical for performance.
4. Out-of-domain data may be useful to improve in-domain performance. (transfer learning philosophy)
References