Learning with Limited Labels for NLP

Applied Deep Learning

YUN-NUNG (VIVIAN) CHEN

May 21st, 2019

HTTP://ADL.MIULAB.TW
Outline

Limited Labeled Data
- How to incorporate the prior knowledge
- How to utilize the current observations

Unlabeled Data
- How to re-use the trained dialogue acts
- How to share knowledge across languages
- How to utilize parallel data

Conclusions
Outline

Limited Labeled Data
  ◦ How to incorporate the prior knowledge: Knowledge-Guided Model
  ◦ How to utilize the current observations

Unlabeled Data
  ◦ How to re-use the trained dialogue acts
  ◦ How to share knowledge across languages
  ◦ How to utilize parallel data

Conclusions
Prior Structural Knowledge

**Sentence s**: show me the flights from seattle to san francisco

**Syntax (Dependency Tree)**

```
  ROOT
    ↓
  show
    ↓
1. me
    ↓
2. the
    ↓
3. seattle
    ↓
4. san
```

**Semantics (AMR Graph)**

```
  show
  ↓
1. you
  ↓
2. flight
  ↓
3. city
  ↓
4. Seattle
  ↓
5. San Francisco
```

Prior knowledge about syntax or semantics may guide understanding

**K-SAN:** Knowledge-Guided Structural Attention Networks

Prior knowledge as a teacher

**Knowledge Encoding Module**

**ROOT**

show me the flights from seattle to san francisco

knowledge-guided structure \( \{x_i\} \)

**Knowledge Encoding**

**CNN\(_{kg}\)**

**Knowledge Attention Distribution**

\( p_i \)

**Encoded Knowledge Representation**

\( m_i \)

**Weighted Sum**

\( \sum \)

**Inner Product**

\( u \)

**Sentence Encoding**

**CNN\(_{in}\)**

**Knowledge-Guided Representation**

\( \Sigma \)

**NN\(_{out}\)**

**RNN Tagger**

slot tagging sequence \( y \)

Sentence Structural Knowledge

Sentence $s$ show me the flights from seattle to san francisco

Syntax (Dependency Tree)

Semantics (AMR Graph)

Knowledge-Guided Substructure $x_i$

1. show me
2. show flights the
3. show flights from seattle
4. show flights to francisco san

Knowledge-Guided Substructure $x_i$

1. show you
2. show flight seattle
3. show flight san francisco
4. show i

The model will pay more attention to more important substructures that may be crucial for slot tagging.

# K-SAN Experiments

<table>
<thead>
<tr>
<th>ATIS Dataset (F1 slot filling)</th>
<th>Small (1/40)</th>
<th>Medium (1/10)</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagger (GRU)</td>
<td>73.83</td>
<td>85.55</td>
<td>93.11</td>
</tr>
<tr>
<td>Encoder-Tagger (GRU)</td>
<td>72.79</td>
<td>88.26</td>
<td>94.75</td>
</tr>
</tbody>
</table>

## K-SAN Experiments

<table>
<thead>
<tr>
<th>ATIS Dataset (F1 slot filling)</th>
<th>Small (1/40)</th>
<th>Medium (1/10)</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagger (GRU)</td>
<td>73.83</td>
<td>85.55</td>
<td>93.11</td>
</tr>
<tr>
<td>Encoder-Tagger (GRU)</td>
<td>72.79</td>
<td>88.26</td>
<td>94.75</td>
</tr>
<tr>
<td>K-SAN (Stanford dep)</td>
<td><strong>74.60⁺</strong></td>
<td>87.99</td>
<td>94.86⁺</td>
</tr>
<tr>
<td>K-SAN (Syntaxnet dep)</td>
<td>74.35⁺</td>
<td><strong>88.40⁺</strong></td>
<td><strong>95.00⁺</strong></td>
</tr>
</tbody>
</table>

*Syntax provides richer knowledge and more general guidance when less training data.*

---

## K-SAN Experiments

<table>
<thead>
<tr>
<th>ATIS Dataset (F1 slot filling)</th>
<th>Small (1/40)</th>
<th>Medium (1/10)</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagger (GRU)</td>
<td>73.83</td>
<td>85.55</td>
<td>93.11</td>
</tr>
<tr>
<td>Encoder-Tagger (GRU)</td>
<td>72.79</td>
<td>88.26</td>
<td>94.75</td>
</tr>
<tr>
<td>K-SAN (Stanford dep)</td>
<td><strong>74.60</strong>+</td>
<td>87.99</td>
<td>94.86+</td>
</tr>
<tr>
<td>K-SAN (Syntaxnet dep)</td>
<td>74.35+</td>
<td><strong>88.40</strong>+</td>
<td><strong>95.00</strong>+</td>
</tr>
<tr>
<td>K-SAN (AMR)</td>
<td>74.32+</td>
<td>88.14</td>
<td>94.85+</td>
</tr>
<tr>
<td>K-SAN (JAMR)</td>
<td>74.27+</td>
<td>88.27+</td>
<td>94.89+</td>
</tr>
</tbody>
</table>

Syntax provides richer knowledge and more general guidance when less training data.

Semantics captures the most salient info so it achieves similar performance with much less substructures.

---

Attention Analysis

Darker blocks and lines correspond to higher attention weights

Attention Analysis

Darker blocks and lines correspond to higher attention weights

Using less training data with K-SAN allows the model to pay the similar attention to the salient substructures that are important for tagging.

EHR Data

Predicting diagnosis codes for clinical reports

- Present illness text
  - “fever up to 39.4°C intermittent in recent 3 days, cough/sputum(+), shortness of breath tonight”

- ICD-9 diagnosis codes
  - 486: Pneumonia, organism unspecified; 780.6: Fever
CNN for Diagnosis Code Prediction
(Li et al., 2017)

Convolutional neural network (CNN) for multi-label code prediction
- Multiple convolutional filters for extracting different patterns

Clinic Text

No dizziness No fever ...

Multilabel Code Prediction

Fully-Connected
Max Pooling
Conv Layer
Embedding Layer

Clinic Text
Hierarchy Category Knowledge

Idea: category knowledge provides additional cues to know code relatedness

Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the “personality disorders”
Hierarchy Category Knowledge (Cluster Penalty)

Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the “personality disorders”

Category constrained loss

\[
\Omega_{\text{between}} = \sum_{k=1}^{K} \| \tilde{\theta}_k - \bar{\theta} \|^2 \\
\Omega_{\text{within}} = \sum_{k=1}^{K} \sum_{i \in \mathcal{I}(k)} \| \theta_i - \tilde{\theta}_k \|^2
\]

Hierarchy Category Knowledge (Multi-Task)

Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the “personality disorders”

Low-level code infers the high-level category

\[ y_{\text{high}} = 1 \text{ if } y_{\text{low}} = 1 \]

Category integrated loss via multi-task

\[ L = L_{\text{low}} + \gamma \cdot L_{\text{high}} \]
Hierarchy Category Knowledge (Avg Meta-Label)

Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the “personality disorders”

High-level prob can be approximated by the average of low-level code prob

\[ y_{high} = \frac{1}{k} \sum y_{low}^k \]

Category integrated loss
\[ L = L_{low} + \gamma \cdot L_{high} \]
Hierarchy Category Knowledge (At-Least-One Meta-Label)

Low-level code
- 301.0: Paranoid personality disorder
- 301.1: Affective personality disorder
- 301.2: Schizoid personality disorder

High-level category
- All belong to the “personality disorders”

High-level prob can be approximated by the at-least-one of low-level code prob

\[
y_{\text{high}} = 1 - \prod_{k} (1 - y_{\text{low}}^k)
\]

Category integrated loss

\[
L = L_{\text{low}} + \gamma \cdot L_{\text{high}}
\]
## State-of-the-Art Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>MIMIC3-50</th>
<th>MIMIC3-Full</th>
<th>CAML (Mullenbach et al., 2018)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P@1</td>
<td>P@3</td>
<td>P@5</td>
</tr>
<tr>
<td>CNN (Shi et al., 2017)</td>
<td>82.8</td>
<td>71.2</td>
<td>61.4</td>
</tr>
<tr>
<td>+ Cluster Penalty</td>
<td>83.5</td>
<td>71.9</td>
<td>62.4</td>
</tr>
<tr>
<td>+ Multi-Task</td>
<td>83.5</td>
<td>71.3</td>
<td>61.9</td>
</tr>
<tr>
<td>+ Hierarchical</td>
<td>84.5</td>
<td>72.1</td>
<td>62.4</td>
</tr>
<tr>
<td>avg at-least-one</td>
<td>83.4</td>
<td>72.1</td>
<td>62.4</td>
</tr>
</tbody>
</table>
Outline

Limited Labeled Data
- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

Unlabeled Data
- How to re-use the trained dialogue acts
- How to share knowledge across languages
- How to utilize parallel data

Conclusions
Semi-Supervised Multi-Task SLU (Lan et al., 2018)

Idea: language understanding objective can enhance other tasks

BLM exploits the *unsupervised knowledge*, the *shared-private framework* and *adversarial training* make the slot tagging model more generalized

Semi-Supervised Multi-Task SLU (Lan et al., 2018)

STM – BLSTM for slot tagging

MTL – multi-task learning for STM and LM, where they share the embedding layer

PSEUDO – train an STM with labeled data, generate labels for unlabeled data, and retrain STM

<table>
<thead>
<tr>
<th>Method</th>
<th>5k</th>
<th>10k</th>
<th>15k</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>STM</td>
<td>67.25</td>
<td>71.04</td>
<td>73.94</td>
<td>76.60</td>
</tr>
<tr>
<td>MTL_e</td>
<td>69.57</td>
<td>73.04</td>
<td>75.00</td>
<td>77.24</td>
</tr>
<tr>
<td>PSEUDO</td>
<td>69.82</td>
<td>72.55</td>
<td>74.80</td>
<td>-</td>
</tr>
<tr>
<td>BSPM</td>
<td>68.46</td>
<td>72.52</td>
<td>75.05</td>
<td>77.52</td>
</tr>
<tr>
<td>BSPM+D^{(w)}</td>
<td>71.55</td>
<td>73.67</td>
<td>74.61</td>
<td>77.42</td>
</tr>
<tr>
<td>BSPM+D^{(s)}</td>
<td>70.99</td>
<td>73.58</td>
<td>74.22</td>
<td>77.24</td>
</tr>
</tbody>
</table>

The model is more efficient when the labeled data is limited and the data for LM is more sufficient.

Outline

Limited Labeled Data
- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

Unlabeled Data
- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- How to share knowledge across languages
- How to utilize parallel data

Conclusions
Zero-Shot Intent Expansion (Chen et al., 2016)

Goal: resolve domain constraint and enable flexible intent expansion for unlabeled domains

Training Data

- `<change_note>`
  - “adjust my note”

- `<change_setting>`
  - “volume turn down”

New Intent

- `<change_calendar>`

CDSSM

“postpone my meeting to five pm”

Intent Embedding

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td></td>
</tr>
</tbody>
</table>

Embedding Generation

| K+1 |
| K+2 |

Same dialogue acts can be shared across domains

Same dialogue acts can be shared across domains

**CDSSM**: Convolutional Deep Structured Semantic Models

Semantic Layer: \( y \)
Semantic Projection Matrix: \( W_s \)
Max Pooling Layer: \( I_m \)
Max Pooling Operation
Convolutional Layer: \( I_c \)
Convolution Matrix: \( W_c \)
Word Hashing Layer: \( I_h \)
Word Hashing Matrix: \( W_h \)
Word Sequence: \( x \)

\[ w_1 \quad w_2 \quad w_3 \quad \ldots \quad w_d \]

\[ \text{Utterance} \quad \text{Intent} \]

\[ U \quad I_1 \quad I_2 \quad \ldots \quad I_n \]

CosSim\((U, I_i)\)

\[ P(I_1 | U) \quad P(I_2 | U) \quad P(I_n | U) \]

\[ P(A | U) = \frac{\exp(CosSim(U, I))}{\sum_{A'} \exp(CosSim(U, I'))} \]

CDSSM maps language usage for the same dialogue acts together

Zero-Shot Intent Expansion (Chen et al., 2016)

The expanded models consider new intents without training samples, and produces better understanding for unseen domains with comparable results for seen domains.

Outline

Limited Labeled Data
- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

Unlabeled Data
- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- How to share knowledge across languages: Zero-Shot Crosslingual SLU
- How to utilize parallel data

Conclusions
Zero-Shot Crosslingual SLU (Upadhyay et al., 2018)

Source language: English (full annotations)
Target language: Hindi (limited annotations)

RT: round trip, FC: from city, TC: to city, DDN: departure day name

(a) English Utterance
Utt: find a one way flight from boston to atlanta on wednesday
Slots: O O B-RT I-RT O O B-FC O B-TC O B-DDN

(b) Hindi Utterance
Utt: बुधवार को बोस्टन से अटलांटा तक जाने वाली एकतरफ़ा उड़ाने खोजें
Slots: B-DDN O B-FC O B-TC O O O B-RT O O
Zero-Shot Crosslingual SLU (Upadhyay et al., 2018)

MT system is not required and both languages can be processed by a single model.

Joint Model for Crosslingual SLU

For rare slots (like meal, airline code), there is a huge difference between the bilingual model and the naive model when the target training data is limited.
Bilingual Model SLU Experiments

The bilingual model outperforms others and does not suffer from latency introduced by MT

Outline

Limited Labeled Data
- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

Unlabeled Data
- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- How to share knowledge across languages: Zero-Shot Crosslingual SLU
- How to utilize parallel data: Crosslingual Sense Embeddings

Conclusions
Crosslingual Embeddings

Tokens in source language shall be mapped to tokens in target language
- This assumption only holds in sense level token
- Sets of crosslingual sense embeddings are therefore important
- uniform/制服 are all polysemous words
Embeddings in a Unified Space (Conneau et al., 2017; Lample et al., 2017)

May largely benefit tasks such as unsupervised machine translation
Our method can be separated into two steps (Lee & Chen, 2017):

1. Select the most probable (argmax) sense given the context
2. Use skip-gram to train the representation of the selected senses

➢ Reinforcement learning is used to connected the two modules
Sense Selection Module

Input:
- Chinese text context \( C_t = [C_{t-m}, ..., C_t = w_i, ..., C_{t+m}] \)
- English text context \( C'_t = [C'_{t-m}, ..., C'_t = w'_i, ..., C'_{t+m}] \)

Output: the fitness for each sense \( z_{i1}, ..., z_{i3} \)

Model architecture: Continuous Bag-of-Words (CBOW) for efficiency

Sense selection

\[
\bar{C} = \alpha \cdot \frac{1}{|c_i|} \sum_{w_i \in c_i} P_{j}^{en} + (1 - \alpha) \cdot \frac{1}{M} \sum_{w'_i \in c'_i} P_{j}^{zh}
\]

\[
p(z_{ik} | c_i, c'_i) = \sigma((Q_{ik})^T \bar{C}')
\]

\[
z_{ik}^* = \arg \max_{z_{ik}} p(z_{ik} | c_i, c'_i)
\]
Sense Representation Module

Input: sense collocation $s_i, s_j, s_l'$

Output: collocation likelihood estimation

Model architecture: skip-gram architecture

Sense selection (optimized by negative sampling)

$$p(s_l' | s_i) = \frac{\exp((U_{s_i}^{en})^T V_{s_l'}^{zh})}{\sum_{s_k'} \exp((U_{s_i}^{en})^T V_{s_k'}^{zh})}$$
Crosslingual Model Architecture

Enabling bilingual sense embedding learning with parallel data
## Qualitative Analysis

<table>
<thead>
<tr>
<th>Target</th>
<th>kNN Senses (EN)</th>
<th>kNN Senses (ZH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>apple_0</td>
<td>fruit, cake, sweet</td>
<td>蘋果, 春天, 蛋糕, iphone, 雞蛋, 巧克力, 葡萄 (apple, spring, cake, iphone, egg, chocolate, purples)</td>
</tr>
<tr>
<td>apple_1</td>
<td>iphone, cake, google, stores</td>
<td>蘋果, iphone, 微軟, 競爭對手, 春天, 谷歌 (apple, iphone, microsoft, competitor, spring, google)</td>
</tr>
<tr>
<td>uniform_0</td>
<td>dressed, worn, tape, wearing, cloth</td>
<td>均勻, 光滑, 衣服, 鞋子, 穿著, 服裝 (even, smooth, clothes, shoes, wearing, clothing)</td>
</tr>
<tr>
<td>uniform_1</td>
<td>particle, computed, varying, gradient</td>
<td>態, 粉末, 縱向, 等離子體, 剪切, 剛度 (phase, powder, longitudinal, plasma, cut, stiffness)</td>
</tr>
</tbody>
</table>

The words with similar senses from both languages have similar embeddings in a unified space
# New Dataset – BCWS
(Bilingual Contextual Word Similarity)

<table>
<thead>
<tr>
<th>English sentence</th>
<th>Chinese sentence</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judges must give both sides an equal opportunity to <code>&lt;state&gt;</code> their cases.</td>
<td>我非常喜歡這個故事，它 <code>&lt;告訴&gt;</code> 我們一些重要的啓示。 (I like this story a lot, which <code>&lt;tells&gt;</code> us some important inspiration.)</td>
<td>7.00</td>
</tr>
<tr>
<td>It was of negligible <code>&lt;importance&gt;</code> prior to 1990, with antiquated weapons and few members.</td>
<td>黃斑部病變的預防及早期治療是相當 <code>&lt;重要&gt;</code> 的。 (The prevention and early treatment of macular lesions is very <code>&lt;important&gt;</code>.)</td>
<td>6.94</td>
</tr>
<tr>
<td>Due to the San Andreas Fault bisecting the hill, one side has <code>&lt;cold&gt;</code> water, the other has hot.</td>
<td>水果攤老闆似乎很意外真有人買這 <code>&lt;冷&gt;</code> 貨，露出「你真內行」的眼神與我聊了幾句。 (The owner of the fruit stall seemed surprised that someone bought this <code>&lt; unpopular&gt;</code> product, talking me few words about “you are such a pro”.)</td>
<td>3.70</td>
</tr>
</tbody>
</table>
The crosslingual sense embeddings learned in an unsupervised way produce better results on BCWS (bilingual) and comparable performance on SCWS (monolingual)
Outline

Limited Labeled Data
- How to incorporate the prior knowledge: Knowledge-Guided Model
- How to utilize the current observations: Semi-Supervised Multi-Task SLU

Unlabeled Data
- How to re-use the trained dialogue acts: Zero-Shot Intent Expansion
- How to share knowledge across languages: Zero-Shot Crosslingual SLU
- How to utilize parallel data: Crosslingual Sense Embeddings

Conclusions
Concluding Remarks

*Prior knowledge* can benefit understanding when less training data

*Language modeling objective* can be incorporated to benefit other tasks

*Dialogue acts* can be shared across different domains

*Crosslingual word embeddings* and *joint model* help extend models to different languages

Sense-level representations can be learned via *contexts*

The *parallel data* for MT can bridge the embeddings from different languages