Semantics as a Foreign Language

Gabriel Stanovsky and Ido Dagan
EMNLP 2018
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms (Oepen et al., 2014; 2015)
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms *(Oepen et al., 2014;2015)*
  - DM (derived from MRS) *(Copestake et al., 1999, Flickinger, 2000)*
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms (Oepen et al., 2014;2015)
  a. DM (derived from MRS)
  b. Prague Semantic Dependencies (PSD) (Hajic et al., 2012)
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms (Oepen et al., 2014; 2015)
  a. DM (derived from MRS)
  b. Prague Semantic Dependencies (PSD)
  c. Predicate Argument Structures (PAS) (Miyao et al., 2014)
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms (Oepen et al., 2014; 2015)
  a. DM (derived from MRS)
  b. Prague Semantic Dependencies (PSD)
  c. Predicate Argument Structures (PAS)

- Aim to capture **semantic** predicate-argument relations
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms *(Oepen et al., 2014;2015)*
  a. DM (derived from MRS)
  b. Prague Semantic Dependencies (PSD)
  c. Predicate Argument Structures (PAS)

- Aim to capture *semantic* predicate-argument relations

- Represented in a graph structure
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms (Oepen et al., 2014; 2015)
  - a. DM (derived from MRS)
  - b. Prague Semantic Dependencies (PSD)
  - c. Predicate Argument Structures (PAS)

- Aim to capture **semantic** predicate-argument relations

- Represented in a graph structure
  - a. Nodes: single words from the sentence
Semantic Dependency Parsing (SDP)

- A collection of three semantic formalisms (Oepen et al., 2014;2015)
  - a. DM (derived from MRS)
  - b. Prague Semantic Dependencies (PSD)
  - c. Predicate Argument Structures (PAS)

- Aim to capture \textbf{semantic} predicate-argument relations

- Represented in a graph structure
  - a. Nodes: single words from the sentence
  - b. Labeled edges: semantic relations, according to the paradigm
Outline
Outline

- SDP as Machine Translation
  - Different formalisms as foreign languages
  - **Motivation**: downstream tasks, inter-task analysis, extendable framework
Outline

- **SDP as Machine Translation**
  - Different formalisms as foreign languages
  - **Motivation**: downstream tasks, inter-task analysis, extendable framework
  - **Previous work** explored the relation between MT and semantics
    - (Wong and Mooney, 2007), (Vinyals et al., 2015), (Flanigan et al., 2016)
Outline

- SDP as Machine Translation
  - Different formalisms as foreign languages
  - **Motivation**: downstream tasks, inter-task analysis, extendable framework
  - **Previous work** explored the relation between MT and semantics
    - (Wong and Mooney, 2007), (Vinyals et al., 2015), (Flanigan et al., 2016)

- Model
  - Seq2Seq
  - Directed graph linearization
Outline

- SDP as Machine Translation
  - Different formalisms as foreign languages
  - **Motivation**: downstream tasks, inter-task analysis, extendable framework
  - **Previous work** explored the relation between MT and semantics
    - (Wong and Mooney, 2007), (Vinyals et al., 2015), (Flanigan et al., 2016)

- Model
  - Seq2Seq
  - Directed graph linearization

- Results
  - Raw text to SDP (near state-of-the-art)
  - Novel inter-task analysis
Outline

● **SDP as Machine Translation**
  ○ Different formalisms as foreign languages
  ○ **Motivation**: downstream tasks, inter-task analysis, extendable framework
  ○ **Previous work** explored the relation between MT and semantics
    
    (Wong and Mooney, 2007), (Vinyals et al., 2015), (Flanigan et al., 2016)

● Model
  ○ Seq2Seq
  ○ Linearization

● Results
  ○ Raw text -> SDP (near state-of-the-art)
  ○ Novel inter-task analysis
Semantic Dependencies as MT

Source

Raw sentence

Target
Semantic Dependencies as MT

Source

Raw sentence

Grammar as a foreign language

Syntax

Target
Semantic Dependencies as MT

- **Source**
  - Raw sentence
  - Syntax
  - SDP

- **Target**
  - Grammar of a foreign language
  - This work

Semantic Dependencies as MT
Semantic Dependencies as MT

- Standard MTL: 3 tasks
Semantic Dependencies as MT

- Standard MTL: 3 tasks

- Inter-task translation (9 tasks)
Outline

● SDP as Machine Translation
  ○ Motivation: downstream tasks
  ○ Different formalisms as foreign languages

● Model
  ○ Seq2Seq
  ○ Linearization

● Results
  ○ Raw text -> SDP (near state-of-the-art)
  ○ Novel inter-task analysis
Our Model I: Raw -> SDP^x

- Seq2Seq translation model:
  - Bi-LSTM encoder-decoder with attention

```
<from: RAW>    <to: DM>  the  cat  sat  on  the  mat
```

```
Linear DM
```

Our Model I: Raw -> SDP^x

- Seq2Seq translation model:
  - Bi-LSTM encoder-decoder with attention

```
<from: RAW> the cat sat on the mat
<to: DM>
```

Special from and to symbols
Our Model Ⅱ: SDP$^y$ -> SDP$^x$

- Seq2Seq translation model:
  - Bi-LSTM encoder-decoder with attention

Special <from: PSD> to <to: DM> symbols

Linear DM

Linear PSD
Our Model

Seq2seq prediction requires a 1:1 linearization function
Linearization: Background

- Previous work used bracketed tree linearization

(Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)
Linearization: Background

- Previous work used bracketed tree linearization
  \( \text{ROOT} \)

(Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)
Linearization: Background

- Previous work used bracketed tree linearization
  \((ROOT (NP \ldots))\)

(Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)
Linearization: Background

- Previous work used bracketed tree linearization

\[(ROOT\ (NP\ (NNP)\ NP\ (VP\ messaged\ (NP\ Alice)\ NP)\ VP)\)\]

(Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)
Linearization: Background

- Previous work used bracketed tree linearization

\[(\text{ROOT} \ (\text{NP} \ (\text{NNP} \ John)) \ (\text{NP} \ alice)) \ (\text{VP} \ messaged)) \text{ROOT}\]

(Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)
Linearization: Background

• Previous work used bracketed tree linearization

\[(ROOT\ (NP\ (NNP\ \text{John}\ )NNP\ )NP\ (VP\ \text{messaged}\ (NP\ \text{Alice}\ )NP\ )VP\ )ROOT\]

(Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)
Linearization: Background

- Previous work used bracketed tree linearization
  
  $(ROOT (NP (NNP John )NNP )NP (VP \text{messaged} (NP Alice )NP )VP )ROOT$

  (Vinyals et al., 2015; Konstas et al., 2017; Buys and Blunsom, 2017)

- Depth-first representation doesn’t directly apply to SDP graphs
  - Non-connected components
  - Re-entrenchies
SDP Linearization (Connectivity)

- **Problem:** No single root from which to start linearization
SDP Linearization (Connectivity)

- **Problem:** No single root from which to start linearization
**SDP Linearization** *(Connectivity)*

- **Problem:** *No single root* from which to start linearization

- **Solution:** Artificial SHIFT edges between non-connected adjacent words
  - All nodes are now reachable from the first word
SDP Linearization (Re-entrancies)

- Re-entrancies require a 1:1 node representation
SDP Linearization (Re-entrancies)

- Re-entrancies require a 1:1 node representation
SDP Linearization (Re-entrenacies)

- Re-entrancies require a 1:1 node representation

(relative index / surface form)
SDP Linearization (Re-entrancies)

- Re-entrancies require a 1:1 node representation

(relative index / surface form)
SDP Linearization (Re-entrancies)

- Re-entrancies require a 1:1 node representation

(relative index / surface form)

0/couch-potato compound +1/jocks
SDP Linearization (Re-entrancies)

- Re-entrancies require a 1:1 node representation

0/couch-potato compound +1/jocks shift +1/watching
SDP Linearization (Re-entrancies)

- Re-entrancies require a 1:1 node representation

0/couch-potato compound +1/jocks shift +1/watching ARG1 -1/jocks
Outline

- SDP as Machine Translation
  - Motivation: downstream tasks
  - Different formalisms as foreign languages

- Model
  - Linearization
  - Dual Encoder-Single decoder Seq2Seq

- Results
  - Raw text -> SDP (near state-of-the-art)
  - Novel inter-task analysis
Experimental Setup

- Train samples per task: 35,657 sentences \((\text{Oepen et al., 2015})\)
  - 9 translation tasks
Experimental Setup

- Train samples per task: 35,657 sentences (Oepen et al., 2015)
  - 9 translation tasks

- Total training samples: **320,913 source-target pairs**
Experimental Setup

- Train samples per task: 35,657 sentences (Oepen et al., 2015)
  - 9 translation tasks

- Total training samples: 320,913 source-target pairs

- Trained in batches between the 9 different tasks
Evaluations: RAW → SDP\(_{(x)}\)

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng et al. (2017a)</td>
<td>90.4</td>
<td>92.7</td>
<td>78.5</td>
<td>87.2</td>
</tr>
<tr>
<td>Single</td>
<td>70.1</td>
<td>73.6</td>
<td>63.6</td>
<td>69.1</td>
</tr>
<tr>
<td>MTL _ PRIMARY</td>
<td>82.4</td>
<td>87.2</td>
<td>71.4</td>
<td>80.3</td>
</tr>
<tr>
<td>MTL _ PRIMARY+AUX</td>
<td>87.5</td>
<td>90.9</td>
<td>80.3</td>
<td>86.2</td>
</tr>
</tbody>
</table>

Labeled F1 score
Evaluations: RAW $\rightarrow$ SDP$_{(x)}$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng et al. (2017a)</td>
<td>90.4</td>
<td>92.7</td>
<td>78.5</td>
<td>87.2</td>
</tr>
<tr>
<td><strong>Single</strong></td>
<td>70.1</td>
<td>73.6</td>
<td>63.6</td>
<td>69.1</td>
</tr>
<tr>
<td>MTL PRIMARY</td>
<td>82.4</td>
<td>87.2</td>
<td>71.4</td>
<td>80.3</td>
</tr>
<tr>
<td>MTL PRIMARY+AUX</td>
<td>87.5</td>
<td>90.9</td>
<td><strong>80.3</strong></td>
<td>86.2</td>
</tr>
</tbody>
</table>

Labeled F1 score
Evaluations: RAW $\rightarrow$ SDP$_{(x)}$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng et al. (2017a)</td>
<td>90.4</td>
<td>92.7</td>
<td>78.5</td>
<td>87.2</td>
</tr>
<tr>
<td>Single</td>
<td>70.1</td>
<td>73.6</td>
<td>63.6</td>
<td>69.1</td>
</tr>
<tr>
<td>MTL PRIMARY</td>
<td>82.4</td>
<td>87.2</td>
<td>71.4</td>
<td>80.3</td>
</tr>
<tr>
<td>MTL PRIMARY+AUX</td>
<td>87.5</td>
<td>90.9</td>
<td>80.3</td>
<td>86.2</td>
</tr>
</tbody>
</table>

Labeled F1 score
Evaluations: RAW $\rightarrow$ SDP$_{(x)}$

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peng et al. (2017a)</td>
<td>90.4</td>
<td>92.7</td>
<td>78.5</td>
<td>87.2</td>
</tr>
<tr>
<td>Single</td>
<td>70.1</td>
<td>73.6</td>
<td>63.6</td>
<td>69.1</td>
</tr>
<tr>
<td>MTL PRIMARY</td>
<td>82.4</td>
<td>87.2</td>
<td>71.4</td>
<td>80.3</td>
</tr>
<tr>
<td>MTL PRIMARY+AUX</td>
<td>87.5</td>
<td>90.9</td>
<td><strong>80.3</strong></td>
<td>86.2</td>
</tr>
</tbody>
</table>

Labeled F1 score
Evaluations: $\text{SDP}_{(a)} \rightarrow \text{SDP}_{(b)}$

<table>
<thead>
<tr>
<th>To \ From</th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td>96.1</td>
<td>92.4</td>
<td>94.3</td>
</tr>
<tr>
<td>PAS</td>
<td>95.7</td>
<td>91.7</td>
<td>93.7</td>
<td></td>
</tr>
<tr>
<td>PSD</td>
<td>89.5</td>
<td>87.6</td>
<td>88.6</td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>92.6</td>
<td>91.9</td>
<td>92.1</td>
<td></td>
</tr>
</tbody>
</table>

**Labeled F1 score**
Evaluations: $\text{SDP}_\text{(a)} \rightarrow \text{SDP}_\text{(b)}$

<table>
<thead>
<tr>
<th>To \ From</th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td>96.1</td>
<td>92.4</td>
<td>94.3</td>
</tr>
<tr>
<td>PAS</td>
<td>95.7</td>
<td></td>
<td>91.7</td>
<td>93.7</td>
</tr>
<tr>
<td>PSD</td>
<td>89.5</td>
<td>87.6</td>
<td></td>
<td>88.6</td>
</tr>
<tr>
<td>Avg.</td>
<td>92.6</td>
<td>91.9</td>
<td>92.1</td>
<td></td>
</tr>
</tbody>
</table>

Labeled F1 score

- Translating **between representations** is easier than parsing from raw text
Evaluations: $\text{SDP}_a \rightarrow \text{SDP}_b$

- Translating **between representations** is easier than parsing from raw text
- Easy to convert between PAS and DM
Evaluations: $SDP_{(a)} \rightarrow SDP_{(b)}$

- Translating **between representations** is easier than parsing from raw text
- Easy to convert between PAS and DM
- PSD is a good input, but relatively hard output

<table>
<thead>
<tr>
<th>To \ From</th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td></td>
<td>96.1</td>
<td><strong>92.4</strong></td>
<td>94.3</td>
</tr>
<tr>
<td>PAS</td>
<td>95.7</td>
<td></td>
<td><strong>91.7</strong></td>
<td>93.7</td>
</tr>
<tr>
<td>PSD</td>
<td><strong>89.5</strong></td>
<td>87.6</td>
<td></td>
<td>88.6</td>
</tr>
<tr>
<td>Avg.</td>
<td><strong>92.6</strong></td>
<td>91.9</td>
<td><strong>92.1</strong></td>
<td></td>
</tr>
</tbody>
</table>

**Labeled F1 score**
Conclusions
Conclusions

- Effective graph linearization for SDP
  - Near state-of-the-art results
Conclusions

● Effective graph linearization for SDP
  ○ Near state-of-the-art results

● Inter-task analysis
  ○ Enabled by the generic seq2seq framework
Conclusions

● Effective graph linearization for SDP
  ○ Near state-of-the-art results

● Inter-task analysis
  ○ Enabled by the generic seq2seq framework

● Future work
  ○ Apply linearizations in downstream tasks (NMT)
  ○ Add more representations (AMR, UD)
Conclusions

● Effective graph linearization for SDP
  ○ Near state-of-the-art results

● Inter-task analysis
  ○ Enabled by the generic seq2seq framework

● Future work
  ○ Apply linearizations in downstream tasks (NMT)
  ○ Add more representations (AMR, UD)

Thanks for listening!
BACKUP SLIDES
Evaluations: Node ordering

- Smaller-first ordering consistently does better across all representations

<table>
<thead>
<tr>
<th></th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>86.1</td>
<td>87.7</td>
<td>78.4</td>
<td>84.1</td>
</tr>
<tr>
<td>Sentence order</td>
<td>87.2</td>
<td>90.3</td>
<td>79.9</td>
<td>85.8</td>
</tr>
<tr>
<td>Closest words</td>
<td>87.5</td>
<td>89.8</td>
<td>79.7</td>
<td>85.8</td>
</tr>
<tr>
<td>Smaller-first</td>
<td>87.9</td>
<td>90.9</td>
<td>80.3</td>
<td>86.2</td>
</tr>
</tbody>
</table>
Semantic Formalisms

● Many formalisms try to represent the meaning of a sentence
  ○ MRS, AMR, PSD, SDP, etc…
Semantic Dependencies as MT

- Syntactic parsing as MT ("Grammar as a foreign language", [Vinyals et. al, 2014])

\[
\text{Jane had a cat} \rightarrow (\text{ROOT} \ (S \ (NP \ Jane)_{NP} \ (VP \ had \ (NP \ a \ cat)_{NP} \ )_{VP} \ )_{S} \ )_{ROOT}
\]

- We aim to do the same for SDP
  - The different formalisms as foreign languages
Semantic Dependencies as MT
Our Model

- **Seq2Seq translation model:**
  - Bi-LSTM encoder-decoder with attention

- **Two shared encoders**
  - From raw to SDP graphs
  - Between SDP graphs

- **One global decoder for all samples**

- **Add “<from:X> <to:Y>” tags to input as preprocessing**
  - Where X, Y in {RAW, PSD, PAS, DM}
  - Different than Google’s NMT, which didn’t have <from:X> tags
    - No “code-switching” is allowed
Motivation

- Linearization is an easy way to **plug-in** predicted structures in NNs
  - MT Target side syntax
    - (Aharoni and Goldberg, 2017; Wang et al., 2018)
- Allows Inter-task analysis
Motivation

- Linearization is an easy way to **plug-in** predicted structures in NNs
  - MT Target side syntax
    (Aharoni and Goldberg, 2017; Wang et al., 2018)
Motivation

- Linearization is an easy way to **plug-in** predicted structures in NNs
  - MT Target side syntax
    (Aharoni and Goldberg, 2017; Wang et al., 2018)

- Allows Inter-task analysis

- Easily extendable framework
SDP Linearization (node ordering)
SDP Linearization (node ordering)
SDP Linearization (node ordering)
SDP Linearization (node ordering)

- Neighbor orderings:
  a. Random - (play, for, jocks, now)
  b. Closest-first
  c. Sentence-order
  d. Smaller-first
SDP Linearization

- Neighbor orderings:
  a. Random
  b. Closest-first - (now, for, play, jocks)
  c. Sentence-order
  d. Smaller-first
SDP Linearization

- Neighbor orderings:
  a. Random
  b. Closest-first
  c. Sentence-order - (jocks, now, for, play)
  d. Smaller-first
SDP Linearization

- Neighbor orderings:
  a. Random
  b. Closest-first
  c. Sentence-order
  d. Smaller-first - (now, play, for, jocks)
Evaluations: Node ordering

- **Smaller-first** ordering consistently does better across all representations

<table>
<thead>
<tr>
<th>Method</th>
<th>DM</th>
<th>PAS</th>
<th>PSD</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>86.1</td>
<td>87.7</td>
<td>78.4</td>
<td>84.1</td>
</tr>
<tr>
<td>Sentence order</td>
<td>87.2</td>
<td>90.3</td>
<td>79.9</td>
<td>85.8</td>
</tr>
<tr>
<td>Closest words</td>
<td>87.5</td>
<td>89.8</td>
<td>79.7</td>
<td>85.8</td>
</tr>
<tr>
<td>Smaller-first</td>
<td><strong>87.9</strong></td>
<td><strong>90.9</strong></td>
<td><strong>80.3</strong></td>
<td><strong>86.2</strong></td>
</tr>
</tbody>
</table>

Labeled F1 score