Semantic Parsing for High-Precision Semantic Role Labelling

Paola Merlo

University of Geneva
Joint work with Gabriele Musillo.
Thanks to James Henderson, Ivan Titov, and the Swiss National Science Foundation.
The Supervised Learning Methodology

LING $\Rightarrow$ Annotation $\Downarrow$

CORPUS $\Rightarrow$ (Quantitative) model of grammar and performance

CS/STAT $\Rightarrow$ Learner $\uparrow\downarrow$

Model $\Rightarrow$ Learnability Studies

Paola Merlo
Semantic Parsing for High-Precision Semantic Role Labelling
Shallow Natural Language Understanding

- **Dialogue**
  - I would like to reserve a flight from Geneva to Boston
  - reserve(THEME=flight, SRC= Geneva, DIR=Boston)

- **Machine Translation**
  - I like it = EXP PRED THEME
  - Mi piace = (THEME) EXP PRED
Different Methods for Semantic Role Labelling

- Rule-based: manipulation of explicit knowledge representation
- Corpus-based + Machine Learning Algorithms
  - Shallow Parsing Pipeline
  - Integrated Full Parsing
Lexical Semantics Annotation: PropBank

The executives gave the chefs a standing ovation
ARG0: the executives
REL: gave
ARG2: the chefs
ARG1: a standing ovation

- Annotation of corpus, representative sampling
- Arguments of the predicate are annotated with **verb-specific** abstract semantic role labels A0-A5 or AA.
- Adjuncts of the predicates are annotated with **general** abstract semantic role labels that are inherited from function labels, such as AM-LOC, AM-TMP, AM-ADV.
the government’s borrowing authority dropped at midnight from $2.80 trillion to $2.87 trillion.
Several current approaches to semantic role labelling perform the labelling in several stages:

1. Prune space of potential arguments
2. Argument identification
3. Argument labelling
Lessons from Previous Work

- Collect statistical indicators of syntactic properties in large tagged corpus to create statistical summary of lexical semantic properties of verbs.
- Use these statistics to classify verbs into classes.
- Classes of verbs can be learnt at approximately 80% accuracy.
- Verb Classification: **Syntax and semantics are highly correlated in corpus statistics**
**Semantic Role Labels Can Be Learned and Recovered While Parsing**

- **Question 1:** can it be done? Task is more difficult and information is less as only part of tree is available.
- **Questions 2:** Can we learn semantic role labels robustly without parsing degradation? Despite increased data sparseness and variability.
- **Question 3:** what are the properties of an integrated approach to semantic role labelling and how can they be exploited at their best?
Developing a model of semantic parsing

- **Syntactic Parsing**: mapping a potentially unbounded string to a tree through a finite set of operations

- **Probabilistic Syntactic Parsing**: Probabilistic parsing handles ambiguities

- **Estimation of Probabilities**: probabilities depend on (potentially unbounded) previous history of derivation

- **Probabilistic Semantic Parsing**: based on linguistically justified correspondence between the syntactic tree and the semantic labels
Syntactic Parsing

- **Syntactic Parsing**: mapping a potentially unbounded string to a tree through a finite set of operations

- We can represent each syntactic structure as an unbounded sequence of basic operations $d_1, d_2 \ldots d_m$, called its derivation
An Example Derivation

0 ROOT
An Example Derivation

\[ d_1 = \text{next}(\text{NNP/Mary}) \]
An Example Derivation

\[ d_2 = \text{project}(\text{NP}) \]
An Example Derivation

\[
\text{ROOT} \quad \text{3} \quad S \\
\text{NP} \quad \text{NNP/Mary} \\
\]

\[d_3 = \text{project}(S)\]
Statistical Syntactic and Semantic Parsing
Linguistically Appropriate Biases
Semantic Parsing

An Example Derivation

\[
\text{d}_4 = \text{next} (\text{VBZ}/\text{runs})
\]

\[
\begin{aligned}
\text{S} & \quad \text{ROOT} \\
\text{NP} & \quad \text{NNP/Mary} \quad 4\text{VBZ}/\text{runs}
\end{aligned}
\]
An Example Derivation

\[ d_5 = \text{project}(\text{VP}) \]

\[
\begin{array}{c}
\text{ROOT} \\
\text{S} \\
\text{NP} \\
\text{NNP/Mary} \\
\text{VP} \\
\text{VBZ/runs}
\end{array}
\]
An Example Derivation

\[ d_6 = \text{next}(\text{RB/often}) \]

```
S
  NP
    NNP/Mary
  VP
    VBZ/runs 6 RB/often
```

Paola Merlo
Semantic Parsing for High-Precision Semantic Role Labelling
An Example Derivation

```
NP  |  ROOT
NNP/Mary  |  VBZ/runs  |  RB/often
```

\[ d_7 = \text{attach} \]
An Example Derivation

ROOT

S

NP

NNP/Mary

VP

VBZ/runs

RB/often

d_8 = attach
An Example Derivation

**ROOT**

- **S**
  - **NP**
    - **NNP/Mary**
  - **VP**
    - **VBZ/runs**
    - **RB/often**

\[ d_9 = \text{attach} \]
Some sentences are ambiguous, so sometimes they have one syntactic structure and sometime they have another.

- *I saw the man with a telescope.*

Many sentences are too complicated for a parser to deduce a single correct structure, because the parser has incomplete information.

The best a parser can do is try to be correct as frequently as possible.

If we guess the **most probable structure**, then we will be correct the most frequently.
How do we know which syntactic structure is the most probable?
History-Based Models

We can estimate the probability of an entire derivation by estimating the probability of each operation conditioned on its history of previous operations

\[
P(d_1, d_2 \ldots d_m) = P(d_1)P(d_2|d_1) \ldots P(d_m|d_1, d_2 \ldots d_{m-1})
\]

\[
= \prod_i P(d_i|d_1, d_2 \ldots d_{i-1})
\]

Now each probability only has a finite number of alternative operations
But these probabilities are still complicated, because the histories can be unbounded in length.

The standard approach is to make independence assumptions, meaning we ignore the less important information in the history.

For example, the context free assumption in Context-Free Grammars.
A Neural Network Statistical Parser

How can we estimate the probabilities $P(d_i|d_1, d_2 \ldots d_{i-1})$ without making independence assumptions (that is without simplifications that might be incorrect)?
A **Simple Synchrony Network** (Henderson 2003) is trained to estimate the probabilities $P(d_i|d_1 \ldots d_{i-1})$.

The SSN first computes a **hidden** representation of the derivation history $h_i = f(d_1 \ldots d_{i-1})$ which compresses the history representation to a finite set of representational units.

The SSN then computes an **output** probability for each possible operation $d_i$ from the hidden representation $h_i$:

$$P(d_i|d_1 \ldots d_{i-1}) \approx P(d_i|h_i)$$
SSN Computation
Learning History Representations

- Any information about $d_1 \ldots d_{i-1}$ input at any earlier step could flow from history representation to history representation and reach the output for $d_i$

- Training is biased toward paying more attention to information which has been input more recently in this flow of information

- We provide an appropriate bias by matching recency in the flow of information between history representations to structural locality between constituents
The Definition of Structurally Local

ROOT

ancestor

9?

self

S

NP

left

child

NNP/Mary

child

VP

VBZ/runs

RB/often

Paola Merlo

Semantic Parsing for High-Precision Semantic Role Labelling
ISBN Parse Example

ROOT

Paola Merlo

Semantic Parsing for High-Precision Semantic Role Labelling
ISBN Parse Example

ROOT

NNP/Mary
ISBN Parse Example

ROOT

NP

NNP/Mary
ISBN Parse Example

ROOT

NP

NNP/Mary

Statistical Syntactic and Semantic Parsing
Linguistically Appropriate Biases
Semantic Parsing
ISBN Parse Example

ROOT
S
NP
NNP/Mary
 VBZ/runs
ISBN Parse Example

```
NP
   NNP/Mary

S
   VP
   VBZ/runs

ROOT
```

Statistical Syntactic and Semantic Parsing
Linguistically Appropriate Biases
Semantic Parsing
ISBN Parse Example

Paola Merlo
Semantic Parsing for High-Precision Semantic Role Labelling
ISBN Parse Example

ROOT

NP

NNP/Mary

VP

VBZ/runs

RB/often
ISBN Parse Example

ROOT

S

NP

NNP/Mary

VP

VBZ/runs

RB/often
If we want to model semantic labels, we need to modify the parser’s structural locality bias. We introduce two kinds of biases to force the parser to pay attention to semantic role labels.
Modelling of Semantic Role Labels

- Lexical biases: Fine-grained modelling of semantic role labelling.
- Network connectivity: highlight portion of tree that bears semantic role labels.
Many successful semantic role labelling systems have tried to model sequences of SRLs.

Semantic roles are mapped onto syntactic structure based on the Thematic Hierarchy
(AG > GOAL/EXP > THEME > DIR/LOC/MNR)
Structural Biases

Regularities, if any, between nodes annotated with semantic labels are captured.
Hypothesis: Semantic role labels are low in tree.

Klein and Manning 03 have shown that a lot can be achieved by splitting tags.

Use SR labels to split ambiguous POS tags.

Lower SR labels (DIR, EXT, LOC, MNR, PNC, CAUS, TMP) onto tags of constituent’s head.
Lowering Semantic Role Labels

The government’s borrowing authority dropped at midnight from $2.80 trillion to $2.87 trillion.
The Experimental Methodology

- SSN model has 613 labels instead of original 33. Vocabulary 4970 tag-word pairs instead of original SSN 512 pairs.
- 240 POS instead of original 45 because of label splitting.
- The baseline: do not modify parser or data in any way to establish level at which task can be performed, if at all.
## Semantic Parsing Results

<table>
<thead>
<tr>
<th></th>
<th>Results on Development Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Baseline</td>
<td>79.6</td>
</tr>
<tr>
<td>Split tags</td>
<td>80.5</td>
</tr>
<tr>
<td>Split tags + enhanced connectivity</td>
<td>81.6</td>
</tr>
</tbody>
</table>
### Indicative Comparison to Pipeline Architectures

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sem-syn SSN</td>
<td>81.6</td>
<td>81.3</td>
<td>81.9</td>
</tr>
<tr>
<td>Best 5 CoNLL</td>
<td>82.7–83.4</td>
<td>82.5–83.1</td>
<td>83.0–83.7</td>
</tr>
</tbody>
</table>
Discussion

- Both enhancements improve over baseline
- Semantic parsing performed almost as well as other pipeline methods which do not produce a rich output
- Syntactic parsing degraded, but still at level of some of the best parsers (88.4% F-measure compared to 88.2% Collins 99)
Application to Semantic Role Labelling

\[ \Rightarrow (v_1, \text{ARG0}, \text{ARG1}, \text{ARG-TMP}) (v_2, \text{ARG0}, \text{ARG1}) \]
Rule-based Extraction Method

- Compile finite state automata gathered from sample of gold data defining path that connect SRL nodes to their predicate
- F-measure of 95.9% on gold data
- F-measure of 69.7% on parser output
SVM Extraction Method

<table>
<thead>
<tr>
<th>verb</th>
<th>role</th>
<th>path</th>
<th>barrier</th>
<th>answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>(v1,</td>
<td>ARG0,</td>
<td>up/up/down,</td>
<td>noS,</td>
<td>yes)</td>
</tr>
<tr>
<td>(v1,</td>
<td>ARG1,</td>
<td>up/down,</td>
<td>noS,</td>
<td>yes)</td>
</tr>
<tr>
<td>(v1,</td>
<td>ARG-TMP,</td>
<td>up/up/down,</td>
<td>noS,</td>
<td>yes)</td>
</tr>
<tr>
<td>(v2,</td>
<td>ARG0,</td>
<td>up/up/up/up/up/down,</td>
<td>S,</td>
<td>no)</td>
</tr>
<tr>
<td>(v2,</td>
<td>ARG0,</td>
<td>up/up/down,</td>
<td>noS,</td>
<td>yes)</td>
</tr>
</tbody>
</table>

SVM classifier

(v1, ARG1, up/down/down, S, ?)
### Results on Development Set (CoNLL Shared Task)

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM All Features</td>
<td>87.4</td>
<td>63.2</td>
<td>73.6</td>
</tr>
<tr>
<td>Rule-based</td>
<td>72.9</td>
<td>66.7</td>
<td>69.7</td>
</tr>
<tr>
<td>SVM No loc/min Features</td>
<td>74.3</td>
<td>63.8</td>
<td>68.6</td>
</tr>
<tr>
<td>Baseline</td>
<td>57.4</td>
<td>53.9</td>
<td>55.6</td>
</tr>
</tbody>
</table>

- **Usefulness of locality features**
Discussion

- Conll shared task, test section
- Results globally not very competitive: bottom third compared to ensemble of learners (79.4–66.7, 75.1)
- Results compared to single systems that do not use any external knowledge: top third (76.4–74.3, 75.1)
- Best precision, low recall (87.6 vs 82.3, 65.8 vs 74.8)
Combination with high-recall systems

- Best precision, low recall (87.6 vs 82.3, 65.8 vs 74.8)
- Combine with method with best recall
- Priority to our system for non-null labels, other labels if we output null
- Results: 80.5% precision, 81.4% recall, 81.0% F-measure
- Better than two systems individually
- Combination of two best systems yields an average between the two (hence not as good as the better of the two)
Current Work

- Improve recall by synchronous assignments (Conll shared task 2008)
- Application to dialogue systems for English and French
Relevance for Linguistics

- Linking theory
- Model of implicit learning and knowledge of language
- Methodology