A Latent Variable Model of Synchronous Parsing for Syntactic and Semantic Dependencies

James Henderson ¹  Paola Merlo ²  Gabriele Musillo ¹ ²  Ivan Titov ³

¹Dept Computer Science, Univ Geneva
²Dept Linguistics, Univ Geneva
³Dept Computer Science, Univ Illinois at U-C

CoNLL 2008
Outline

1. A Latent Variable Model of Synchronous Parsing
2. Probability Model
3. Machine Learning Method
4. Evaluation
Motivation for synchronous parsing

- Syntax and semantics are **separate structures**, with different generalisations
  - Sub
    - John
    - The vase
  - Obj
    - broke
    - the vase.
  - A0
  - A1

- Syntax and semantics are **highly correlated**, and therefore should be learned jointly

- **Synchronous parsing** provides a single joint model of two separate structures
The correlations between syntax and semantics are partly **lexical**, and independence assumptions are **hard to specify** a priori.

The dataset is new, and there was little time for feature engineering.

**Latent variables** provide a powerful mechanism for discovering correlations both within and between the structures.
Outline

1. A Latent Variable Model of Synchronous Parsing
2. Probability Model
3. Machine Learning Method
4. Evaluation
Outline

1. A Latent Variable Model of Synchronous Parsing
2. Probability Model
3. Machine Learning Method
4. Evaluation
The Probability Model

- A **generative, history-based** model
- of the **joint probability**
- of syntactic and semantic **synchronous derivations**
- synchronised at **each word**.
Syntactic and semantic dependencies example

ROOT Hope seems doomed to failure

\[ P(T_d, T_s) \]
Syntactic and semantic derivations

Define **two separate derivations**, one for the syntactic structure and one for the semantic structure.

\[ P(T_d, T_s) = P(D^1_d, \ldots, D^m_d, D^1_s, \ldots, D^m_s) \]

- Actions of an incremental shift-reduce style parser similar to MALT [Nivre et al., 2006]
- Semantic derivations are less constrained, because their structures are less constrained
- Assumes each dependency structure is **individually planar** (“projective”)
Use an intermediate synchronisation granularity, between full predications and individual actions.

\[ C^t = D^d_{b_t}, ..., D^d_{e_t}, \text{shift}_t, D^s_{b_t}, ..., D^s_{e_t}, \text{shift}_t \]

\[ P(D^1_d, ..., D^{md}_d, D^1_s, ..., D^{ms}_s) = P(C^1, \ldots, C^n) \]

- Synchronisation at each word prediction
- Results in one shared input queue
- Allows two separate stacks
Synchronous parsing example

ROOT Hope

\( P(C^1) \)
A Latent Variable Model of Synchronous Parsing
Probability Model
Machine Learning Method
Evaluation

Synchronous parsing example

ROOT  Hope  seems

\[ P(C^1) \cdot P(C^2|C^1) \]
A Latent Variable Model of Synchronous Parsing
Probability Model
Machine Learning Method
Evaluation

Synchronous parsing example

ROOT Hope seems doomed

\[ P(C^1) \ P(C^2|C^1) \ P(C^3|C^1, C^2) \]
Synchronous parsing example

\[ P(C^1) \cdot P(C^2|C^1) \cdot P(C^3|C^1, C^2) \cdot P(C^4|C^1, C^2, C^3) \]
Synchronous parsing example

\[ P(C^1) P(C^2|C^1) P(C^3|C^1, C^2) P(C^4|C^1, C^2, C^3) P(C^5|C^1, C^2, C^3, C^4) \]
Derivation example

ROOT  Hope
Derivation example

ROOT  Hope  seems
Derivation example

ROOT  Hope  seems
Derivation example

ROOT Hope seems
Derivation example

ROOT Hope seems
Derivation example

ROOT  Hope  seems  doomed
Derivation example

ROOT Hope seems doomed
Derivation example

ROOT  Hope  seems  doomed
Derivation example

ROOT Hope seems doomed
Derivation example

ROOT  Hope  seems  doomed  to
Hope seems doomed to
Derivation example

ROOT Hope seems doomed to
Derivation example

ROOT Hope seems doomed to failure
Derivation example

ROOT Hope seems doomed to failure
Derivation example

ROOT Hope seems doomed to failure
Projectivisation

- Allows crossing links **between syntax and semantics**
- Use the HEAD method [Nivre et al., 2006] to projectivise syntax
- Use syntactic dependencies to projectivise semantic dependencies
Projectivising semantic dependencies

A Latent Variable Model of Synchronous Parsing
Probability Model
Machine Learning Method
Evaluation
Outline

1. A Latent Variable Model of Synchronous Parsing
2. Probability Model
3. Machine Learning Method
4. Evaluation
Synchronous derivations are modeled with an Incremental Sigmoid Belief Network (ISBN).

- ISBNs are Dynamic Bayesian Networks for modeling structures,
- with vectors of latent variables annotating derivation states
- that represent features of the derivation history.
Connections between latent states reflect locality in the syntactic or semantic structure, thereby specifying the domain of locality for conditioning decisions. Explicit conditioning features of the history are also specified.
Connections between latent states

- Distinguish between syntactic states and semantic states of the derivation
- Connections both within and between types of states

<table>
<thead>
<tr>
<th>Recent</th>
<th>Current</th>
<th>Syn-Syn</th>
<th>Srl-Srl</th>
<th>Syn-Srl</th>
<th>Srl-Syn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next</td>
<td>Next</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
</tr>
<tr>
<td>Top</td>
<td>Top</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
</tr>
<tr>
<td>RgtDepTop</td>
<td>Top</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LftDepTop</td>
<td>Top</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HeadTop</td>
<td>Top</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LftDepNext</td>
<td>Top</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next</td>
<td>Top</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Explicit conditioning features

<table>
<thead>
<tr>
<th>State</th>
<th>Syntax</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LEX POS DEP</td>
<td>LEX POS DEP SENSE</td>
</tr>
<tr>
<td>Next</td>
<td>+ +</td>
<td>+ + +</td>
</tr>
<tr>
<td>SynTop</td>
<td>+ +</td>
<td>+ + +</td>
</tr>
<tr>
<td>SynTop - 1</td>
<td>+</td>
<td>+ + +</td>
</tr>
<tr>
<td>Head SynTop</td>
<td>+</td>
<td>+ + +</td>
</tr>
<tr>
<td>RgtD SynTop</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>LftD SynTop</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>LftD Next</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

State:
- Next
- SemTop
- RgtD SemTop
- LftD SemTop
- LftD Next
- A0-A5 SemTop
- A0-A5 Next
Outline

1. A Latent Variable Model of Synchronous Parsing
2. Probability Model
3. Machine Learning Method
4. Evaluation
The Evaluation

Two models reported

Submitted model:
- vocabulary of 1083 words
- latent vector of 60 features
- no semantics-to-syntax latent state connections
- a form of Minimum Bayes Risk (MBR) decoding for syntax

Larger model:
- vocabulary of 4392 words
- latent vector of 80 features
- includes semantics-to-syntax latent state connections
- decoding optimises joint probability
## Results

<table>
<thead>
<tr>
<th></th>
<th>Syntactic LAS</th>
<th>Semantic P</th>
<th>Semantic R</th>
<th>Semantic F1</th>
<th>Overall F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Submitted</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSJ</td>
<td>87.8</td>
<td>79.6</td>
<td>66.2</td>
<td>72.3</td>
<td>80.2</td>
</tr>
<tr>
<td>Brn</td>
<td>80.0</td>
<td>66.6</td>
<td>55.3</td>
<td>60.4</td>
<td>70.3</td>
</tr>
<tr>
<td>WSJ+Brn</td>
<td>86.9</td>
<td>78.2</td>
<td>65.0</td>
<td>71.0</td>
<td><strong>79.1</strong></td>
</tr>
<tr>
<td><strong>Large</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSJ</td>
<td>88.5</td>
<td>80.4</td>
<td>69.2</td>
<td>74.4</td>
<td>81.5</td>
</tr>
<tr>
<td>Brn</td>
<td>81.0</td>
<td>68.3</td>
<td>57.7</td>
<td>62.6</td>
<td>71.9</td>
</tr>
<tr>
<td>WSJ+Brn</td>
<td>87.6</td>
<td>79.1</td>
<td>67.9</td>
<td>73.1</td>
<td><strong>80.5</strong></td>
</tr>
</tbody>
</table>

- Larger model does better (1.5%) than smaller submitted model
- Large model would be **fifth** overall
### MBR versus joint inference

<table>
<thead>
<tr>
<th></th>
<th>Syntactic LAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submitted</td>
<td></td>
</tr>
<tr>
<td>Dev</td>
<td><strong>86.1</strong></td>
</tr>
<tr>
<td>Joint optimisation</td>
<td></td>
</tr>
<tr>
<td>Dev</td>
<td><strong>85.5</strong></td>
</tr>
<tr>
<td>Large (joint optimisation)</td>
<td></td>
</tr>
<tr>
<td>Dev</td>
<td><strong>86.5</strong></td>
</tr>
</tbody>
</table>

- MBR for syntax helps a bit (0.6%)
- but not as much as the large model (1.0%)
Additional experiments

- Removing latent connections **between syntax and semantics** reduced semantic performance by 3.5%, indicating the **importance of the latent variables** for finding the correlations between these structures.

- When evaluated only on **syntactic dependencies**, the submitted model performs slightly (0.2%) **better** than a model trained **only on syntactic dependencies**, indicating that training a joint model does not harm performance of the syntax component, and may help.
Conclusions

- Synchronous derivations are an effective way to build joint models of separate structures.
- The latent features of ISBNs help find correlations between structures.
- ISBNs extend well to more complex automata than push-down automata.
Current Work

- Derivations which *projectivise on-line* (81.8% overall F-measure, 1.3% improvement)
- Better feature engineering, particularly for semantic parse decisions
Acknowledgements

This work was partly funded by

- European Community FP7 project CLASSiC (www.classic-project.org),
- a Swiss NSF grant,
- two Swiss NSF fellowships.

Part of this work was done when G. Musillo was visiting MIT/CSAIL.
An arc is un-crossed by replacing its argument $a$ with $a$'s syntactic head and noting this change in the arc label. This change is repeated as necessary using a heuristic greedy search.

Projectivising semantic dependencies
Decoding

- Beam search used to search for the most probable derivation
- For submitted model, chose syntactic structure by summing over beam of semantic structures