# 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]

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

CoNLL 2008

# Outline

# Motivation for synchronous parsing

- Syntax and semantics are **separate structures**, with different generalisations

| | | |
|---|---|---|
| Sub | | Obj |
| John | broke | the vase. |
| A0 | | A1 |

| | |
|---|---|
| Sub | |
| The vase | broke. |
| A1 | |

- Syntax and semantics are **highly correlated**, and therefore should be learned jointly
- **Synchronous parsing** provides a single joint model of two separate structures

# Motivation for latent variables

- 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

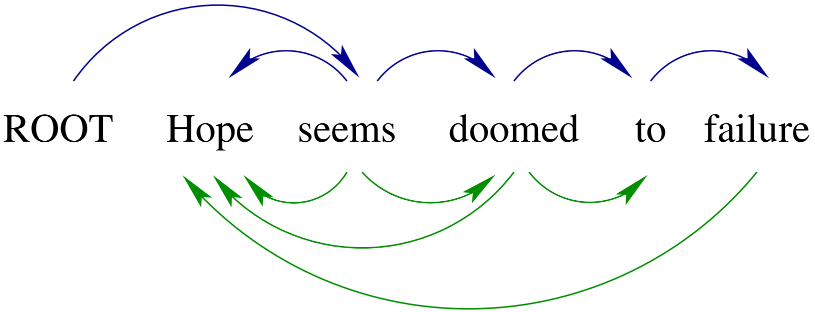
# Outline

# Outline

# 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




$$P(T_d, T_s)$$

# Syntactic and semantic derivations

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

$$P(T_d, T_s) = P(D_d^1, ..., D_d^{m_d}, D_s^1, ..., D_s^{m_s})$$

- 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”)

# Synchronisation granularity

Use an intermediate synchronisation granularity, between full predications and individual actions.

$$C^t = D_d^{b_d^t}, ..., D_d^{e_d^t}, shift_t, D_s^{b_s^t}, ..., D_s^{e_s^t}, shift_t$$

$$P(D_d^1, ..., D_d^{m_d}, D_s^1, ..., D_s^{m_s}) = P(C^1, \ldots, C^n)$$

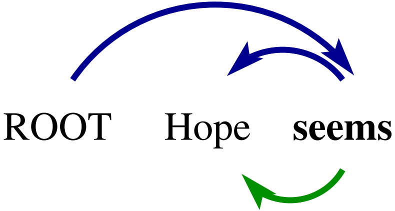
- Synchronisation at **each word** prediction
- Results in **one shared input queue**
- Allows **two separate stacks**

# Synchronous parsing example

ROOT **Hope**

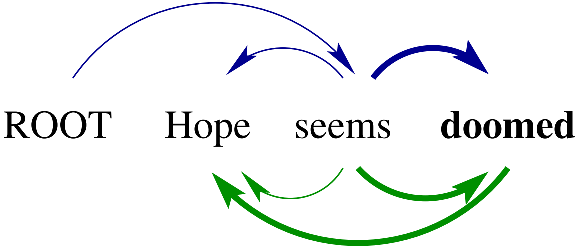
$\mathbf{P}(\mathbf{C}^1)$

# Synchronous parsing example



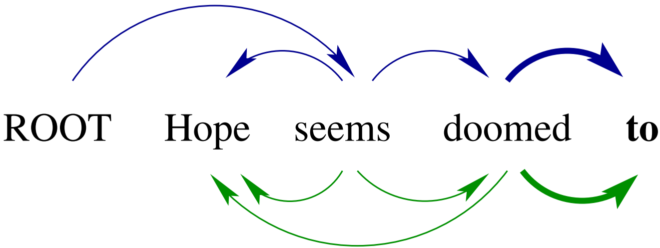

$$P(C^1)\,\mathbf{P}(\mathbf{C^2}|\mathbf{C^1})$$

# Synchronous parsing example



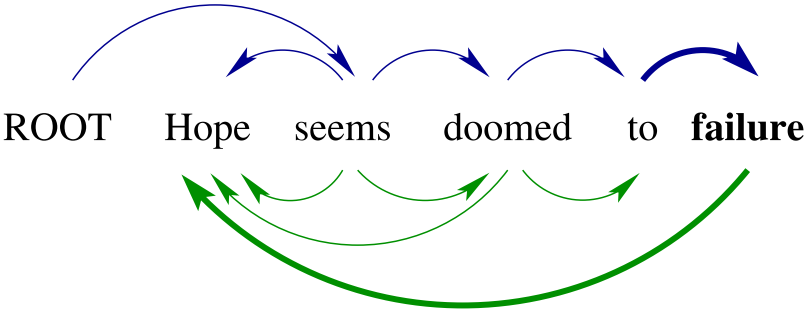

$$P(C^1)\ P(C^2|C^1)\ \mathbf{P}(\mathbf{C^3}|\mathbf{C^1},\mathbf{C^2})$$

# Synchronous parsing example




$$P(C^1)\ P(C^2|C^1)\ P(C^3|C^1, C^2)\ \mathbf{P(C^4|C^1, C^2, C^3)}$$

# Synchronous parsing example




$$P(C^1)\ P(C^2|C^1)\ P(C^3|C^1,C^2)\ P(C^4|C^1,C^2,C^3)\ \mathbf{P(C^5|C^1,C^2,C^3,C^4)}$$

# Derivation example

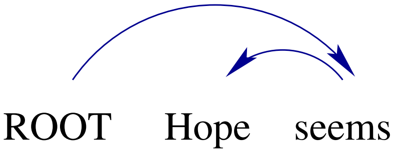
ROOT Hope

# Derivation example

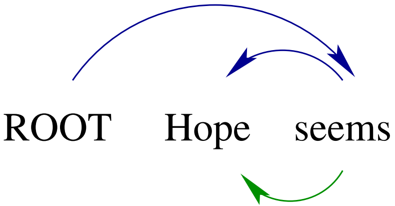
ROOT Hope seems

# Derivation example

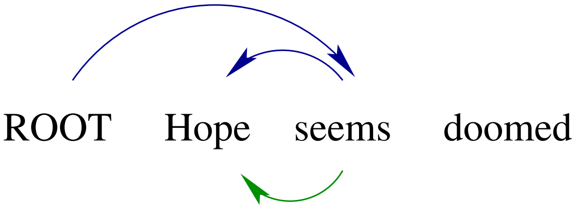
# Derivation example



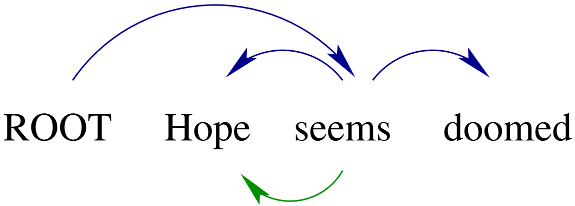
# Derivation example



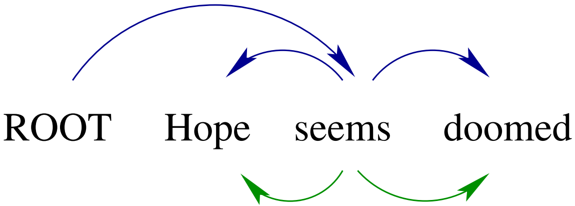
# Derivation example



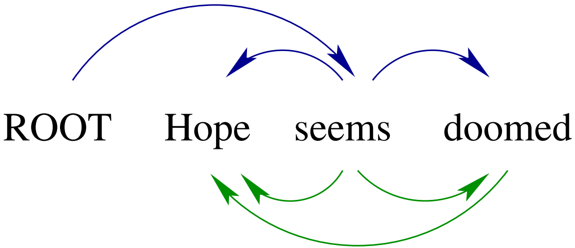
# Derivation example



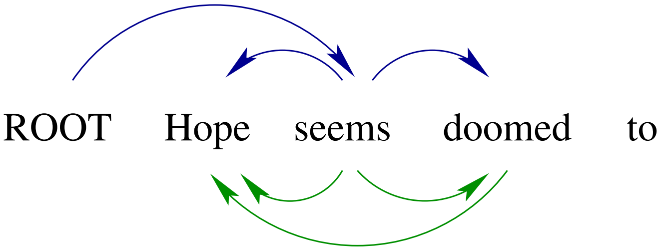
# Derivation example



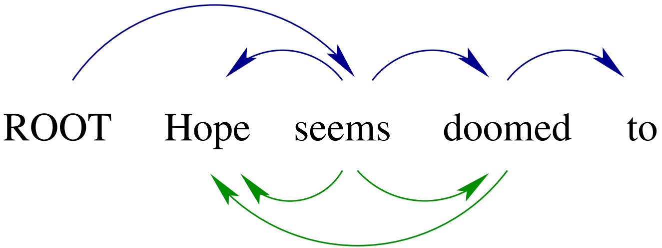
# Derivation example



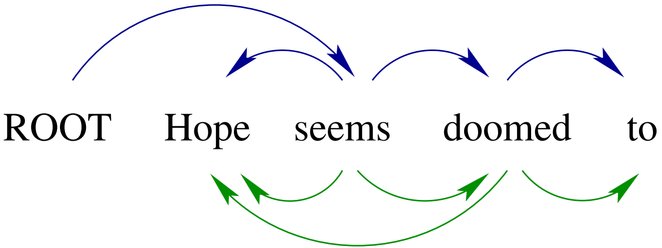
# Derivation example



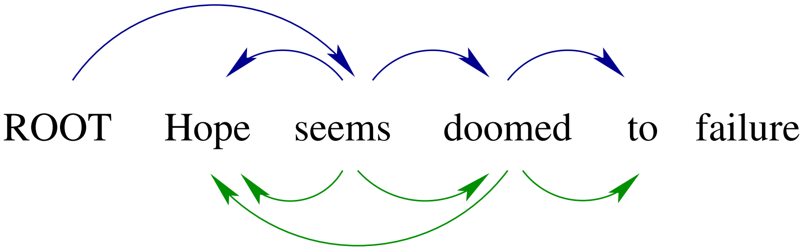
# Derivation example



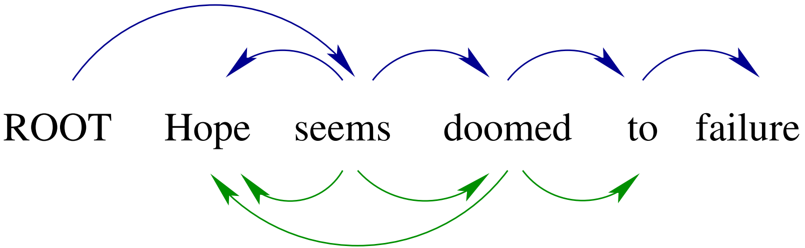
# Derivation example



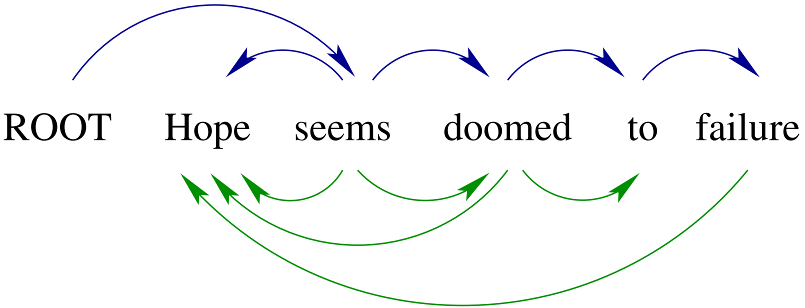
# Derivation example



# Derivation example



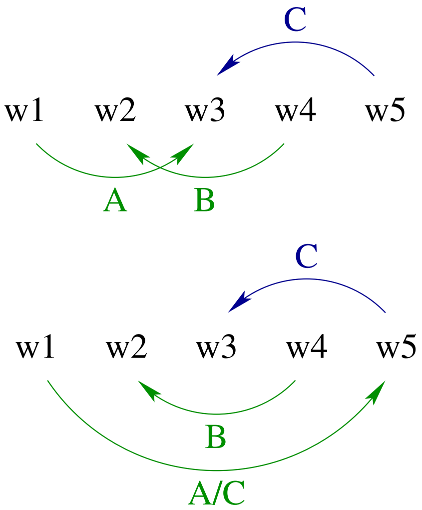
## Derivation example



# 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



# Outline

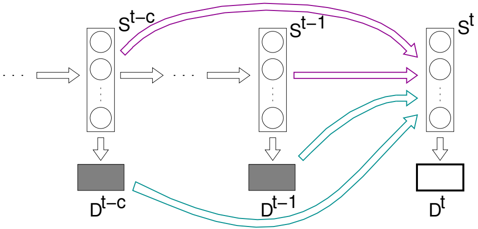
# The Machine Learning Method

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.**
- Use the neural network approximation of ISBNs [Titov and Henderson, ACL 2007] (“Simple Synchrony Netowrks”)

# Statistical dependencies in the ISBN

- 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

| Recent | Current | Syn-Syn | Srl-Srl | Syn-Srl | Srl-Syn |
|---|---|---|---|---|---|
| Next | Next | + | + | + | (+) |
| Top | Top | + | + | + | (+) |
| RgtDepTop | Top | + | + | | |
| LftDepTop | Top | + | + | | |
| HeadTop | Top | + | + | | |
| LftDepNext | Top | + | + | | |
| Next | Top | + | | | |

# Explicit conditioning features

| State | Syntax | | |
|---|---|---|---|
| | LEX | POS | DEP |
| Next | + | + | |
| SynTop | + | + | |
| SynTop - 1 | | + | |
| Head SynTop | + | | |
| RgtD SynTop | | | + |
| LftD SynTop | | | + |
| LftD Next | | | + |

| State | **Semantics** | | | |
|---|---|---|---|---|
| | LEX | POS | DEP | SENSE |
| Next | + | + | | + |
| SemTop | + | + | | + |
| SemTop - 1 | + | + | | |
| Head SemTop | + | | + | |
| RgtD SemTop | | | + | |
| LftD SemTop | | | + | |
| LftD Next | | | + | |
| A0-A5 SemTop | | + | | |
| A0-A5 Next | | + | | |

# Outline

# 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

| | Syntactic | Semantic | | | Overall |
|---|---|---|---|---|---|
| | LAS | P | R | F1 | F1 |
| Submitted | | | | | |
| WSJ | 87.8 | 79.6 | 66.2 | 72.3 | 80.2 |
| Brn | 80.0 | 66.6 | 55.3 | 60.4 | 70.3 |
| WSJ+Brn | 86.9 | 78.2 | 65.0 | 71.0 | **79.1** |
| Large | | | | | |
| WSJ | 88.5 | 80.4 | 69.2 | 74.4 | 81.5 |
| Brn | 81.0 | 68.3 | 57.7 | 62.6 | 71.9 |
| WSJ+Brn | 87.6 | 79.1 | 67.9 | 73.1 | **80.5** |

- Larger model does better (1.5%) than smaller submitted model
- Large model would be **fifth** overall

# MBR versus joint inference

| | Syntactic LAS |
|---|---|
| Submitted | |
| Dev | **86.1** |
| Joint optimisation | |
| Dev | **85.5** |
| Large (joint optimisation) | |
| Dev | 86.5 |

- 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 slighly (0.2%) **better** than a model trained **only on syntactic depedencies**, 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.

# Projectivising semantic dependencies

- 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.

# Decoding

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