UPPSALA UNIVERSITET

# Neural Network Techniques in Dependency Parsing

Joakim Nivre

Uppsala University
Linguistics and Philology

# Overall Plan

1. Basic notions of dependency grammar and dependency parsing
2. Graph-based and transition-based dependency parsing
3. Advanced graph-based parsing techniques
4. Advanced transition-based parsing techniques
5. Neural network techniques in dependency parsing
6. Multilingual parsing from raw text to universal dependencies

# Taking Stock

LAS: 83.8 v. 83.6
[McDonald & Nivre 2007]

| | Graph-based Parsers | Transition-based Parsers |
|---|---|---|
| 2007 | Global Inference<br>Global Learning<br>Local Feature Scope | Local Inference<br>Local Learning<br>Global Feature Scope |
| | ↓ higher-order chart parsing<br>pruning<br>ILP<br>dual decomp<br>mildly non-projective<br>etc. | ↓ beam search<br>perceptron<br>dynamic oracles<br>dynamic programming<br>more features<br>etc. |
| 2014 | **Graph-based Parsers**<br>Global Inference<br>Global Learning<br>Global Feature Scope | **Transition-based Parsers**<br>Global Inference<br>Global Learning<br>Global Feature Scope |

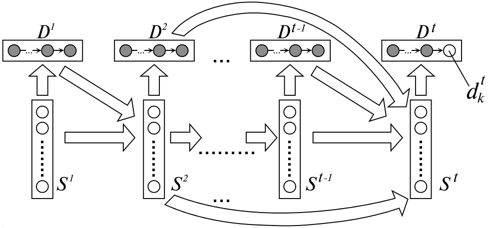
LAS: 85.8 v. 85.5
[Zhang et al. 2013]

**Evaluated on overlapping 9 languages in studies**

# Neural Network Techniques

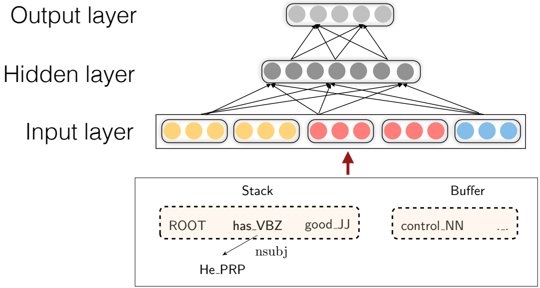
- Empirical results have improved substantially since 2014
- Neural networks techniques yield more effective features:
  - Features are learned (not hand-crafted)
  - Features are continuous and dense (not discrete and sparse)
  - Features can be tuned to (multiple) specific tasks
  - Features can capture unbounded dependencies
- Parsing architectures remain essentially the same

# Learning Features [Titov and Henderson 2007]



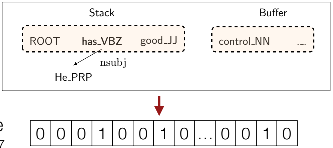

- Incremental Sigmoid Belief Network (ISBN)
- Learns complex features using binary latent variables
- Captures dependencies at arbitrarily long distances
- First generative model for transition-based parsing

# Learning Dense Features [Chen and Manning 2014]




- MaltParser with MLP instead of SVM (greedy, local)
- But 2 percentage points better LAS on PTB/CTB!?

# Traditional Features [Chen and Manning 2014]



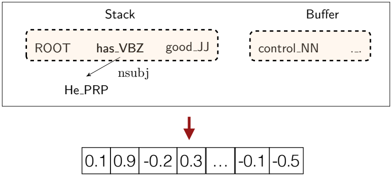

binary, sparse
dim $=10^6 \sim 10^7$

Indicator features

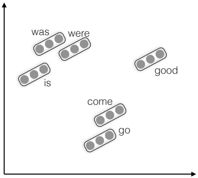
$$s_2.w = \text{has} \wedge s_2.t = \text{VBZ}$$
$$s_1.w = \text{good} \wedge s_1.t = \text{JJ} \wedge b_1.w = \text{control}$$
$$lc(s_2).t = \text{PRP} \wedge s_2.t = \text{VBZ} \wedge s_1.t = \text{JJ}$$
$$lc(s_2).w = \text{He} \wedge lc(s_2).l = \text{nsubj} \wedge s_2.w = \text{has}$$

- Sparse – but lexical features and interaction features crucial
- Incomplete – unavoidable with hand-crafted feature templates
- Expensive – accounts for 95% of computing time

# Dense Features [Chen and Manning 2014]

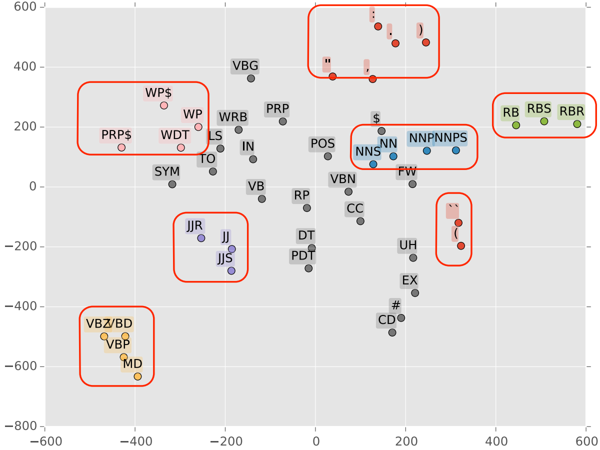



dense
dim = 200

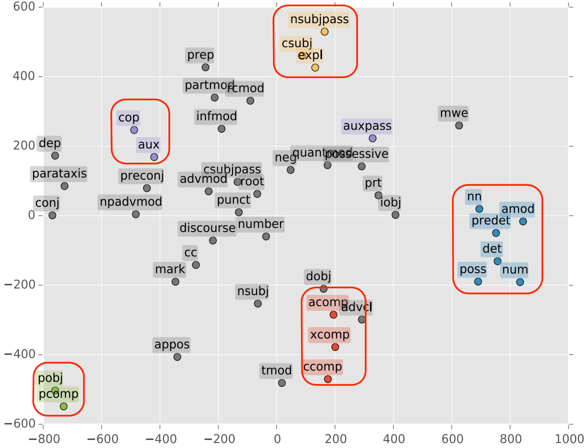



- Sparse – dense features capture similarities (words, pos, dep)
- Incomplete – neural network learns interaction features
- Expensive – matrix multiplication with low dimensionality

# PoS Embeddings [Chen and Manning 2014]

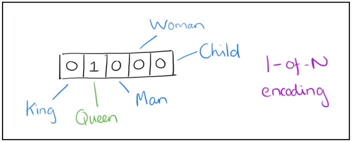


# Dep Embeddings [Chen and Manning 2014]

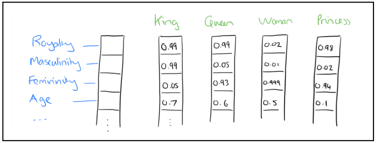


# The Power of Embeddings

**One-Hot** (discrete, sparse)




**Embedding** (continuous, dense)

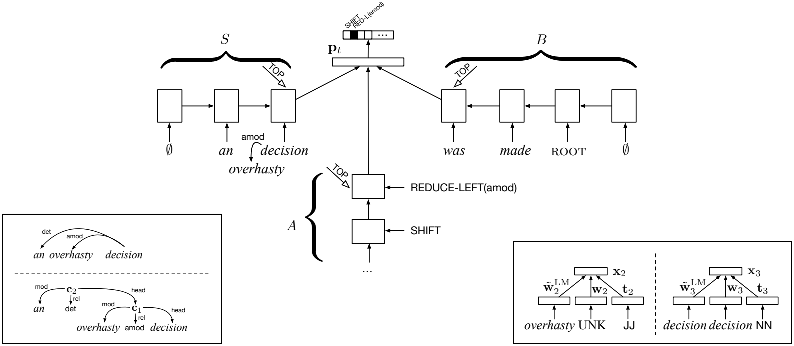



- Inherently much more expressive ($\mathcal{R} \times D$ vs. 1)
- Can capture similarities between items (sparsity)
- Can be pre-trained on large unlabeled corpora (OOV)
- Can be learned/tuned specifically for the parsing task

# Neural Dependency Parsing

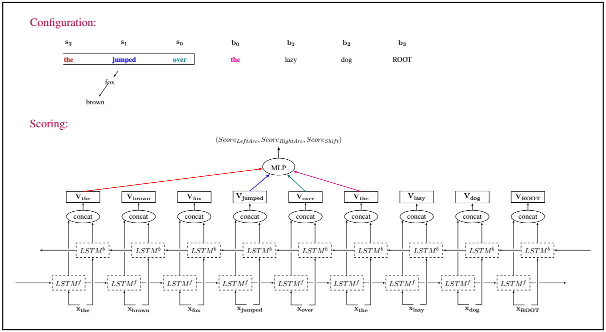
- Dominated by transition-based approaches
- Two main lines of work:
  - More powerful (recurrent) neural networks
    [Dyer et al. 2015, Kiperwasser and Goldberg 2016]
  - Global optimization and beam search
    [Weiss et al. 2015, Andor et al. 2016]
- Additional themes:
  - Character-based models for morphologically rich languages
    [Ballesteros et al. 2015]
  - Cross-lingual embeddings and typological features
    [Ammar et al. 2016]

# Stack-LSTM [Dyer et al. 2015]




- LSTM encoding of parser configurations (S, B, A)
- Stack elements recursively composed of word representations

# Bi-LSTM [Kiperwasser and Goldberg 2016]




- Bi-LSTM encodes global context in word representations
- Exploration with dynamic oracles prevent error propagation

# Global Normalization [Andor et al. 2016]

**Local**

$$p_L(d_{1:n}) = \prod_{j=1}^{n} p(d_j | d_{1:j-1}; \theta)$$

$$p(d_j | d_{1:j-1}; \theta) = \frac{\exp \rho(d_{1:j-1}, d_j; \theta)}{Z_L(d_{1:j-1}; \theta)}$$

$$Z_L(d_{1:j-1}; \theta) = \sum_{d' \in \mathcal{A}(d_{1:j-1})} \exp \rho(d_{1:j-1}, d'; \theta)$$

**Global**

$$p_G(d_{1:n}) = \frac{\exp \sum_{j=1}^{n} \rho(d_{1:j-1}, d_j; \theta)}{Z_G(\theta)}$$

$$Z_G(\theta) = \sum_{d'_{1:n} \in \mathcal{D}_n} \exp \sum_{j=1}^{n} \rho(d'_{1:j-1}, d'_j; \theta)$$

- Global normalization $\rightarrow$ sum over all transition sequences
- Approximation using beam search and early update

# Evaluation

| System | UAS | LAS | Approach |
|---|---|---|---|
| Zhang and Nivre (2011) | 93.5 | 91.9 | Transition, struct perc, beam |
| Martins et al. (2013) | 92.9 | 90.6 | Graph, 3rd-order, dual decomp |
| Zhang and McDonald (2014) | 92.9 | 90.6 | Graph, 3rd-order, cube pruning |
| Dyer et al. (2015) | 93.1 | 90.9 | Transition, LSTM, greedy |
| Kiperwasser et al. (2016) | 93.9 | 91.9 | Transition, LSTM/MLP, greedy |
| Weiss et al. (2015) | 94.0 | 92.0 | Transition, MLP, beam |
| Andor et al. (2016) | 94.6 | 92.8 | Transition, MLP global, beam |

Evaluation on WSJ with Stanford Dependencies

# Taking Stock Again

- Traditional architectures persist
  - When will we see a new dependency parsing algorithm?
  - Do we even need parsing algorithms?
- Transition-based parsers dominate
  - Rich features trump global learning/inference?
  - Or will the empire strike back?
- Predicting the future is hard . . .

# Coming Up Next

1. Basic notions of dependency grammar and dependency parsing
2. Graph-based and transition-based dependency parsing
3. Advanced graph-based parsing techniques
4. Advanced transition-based parsing techniques
5. Neural network techniques in dependency parsing
6. Multilingual parsing from raw text to universal dependencies

## References and Further Reading

- Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016.
  Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.
- Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016.
  Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452.
- Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015.
  Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 349–359.
- Danqi Chen and Christopher Manning. 2014.
  A fast and accurate dependency parser using neural networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 740–750.
- Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015.

Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 334–343.

- Eliyahu Kiperwasser and Yoav Goldberg. 2016.
  Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

- Ivan Titov and James Henderson. 2007.
  A latent variable model for generative dependency parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*, pages 144–155.

- David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015.
  Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 323–333.