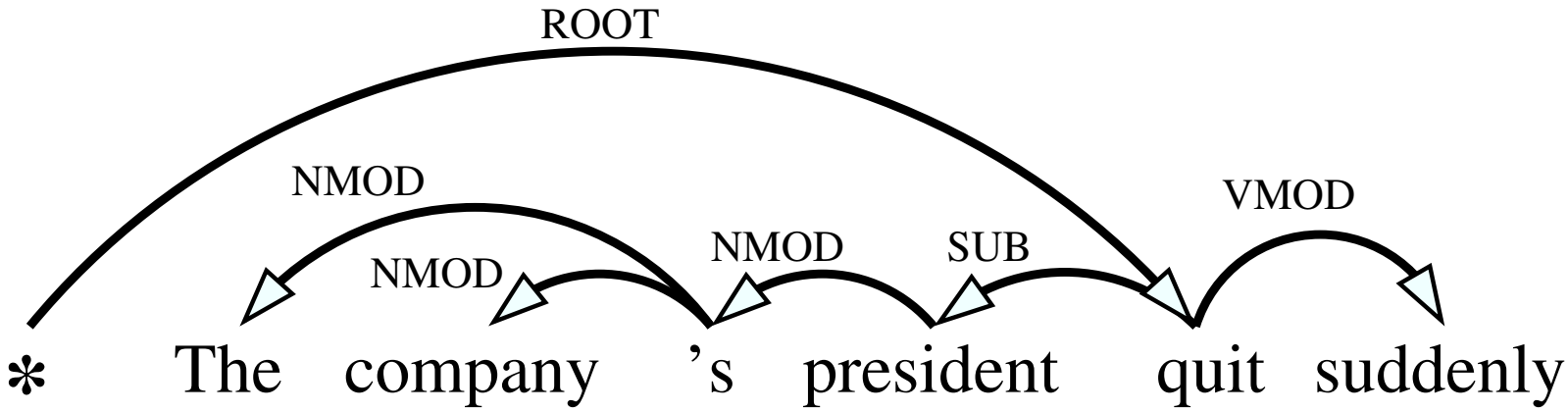
Simple Semi-supervised Dependency Parsing

Terry Koo, Xavier Carreras and Michael Collins

{maestro,carreras,mcollins}@csail.mit.edu

Dependency Parsing



- Statistical parsers make heavy use of lexicalized statistics, which are sparse
- Alternate lexical representations?

Our Approach

- Derive lexical representations (hierarchical word clusterings) from unlabeled data
- Clusters are easily incorporated as features in a discriminative dependency parser
- Improvements over state-of-the-art baselines in English and Czech

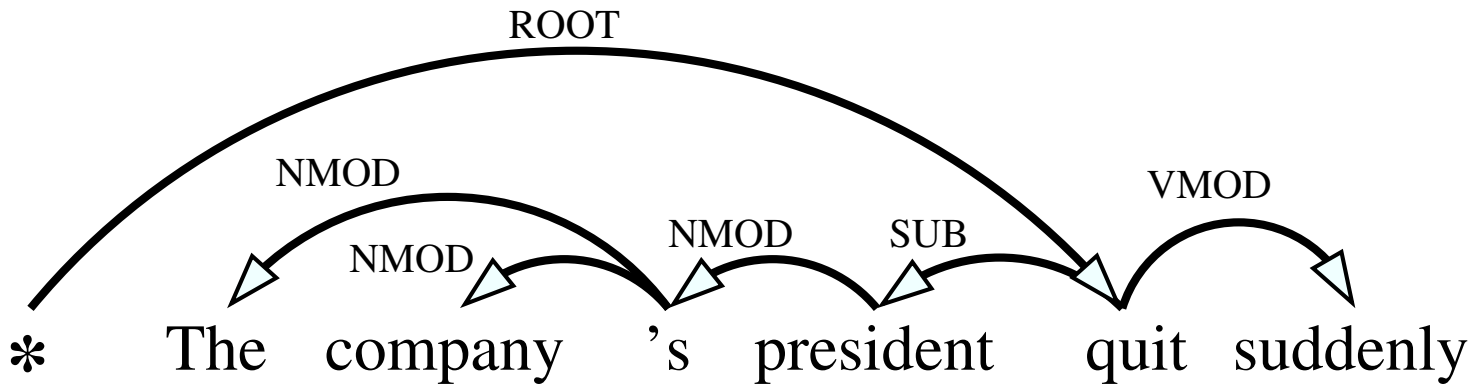
Previous Work

- Brown et al. (1992): clustering algorithm, applied to language modeling
- Miller et al. (2004): named-entity tagging with word clusters from the Brown algorithm
- McDonald et al. (2005, 2006): discriminative dependency parsing with flexible feature-vector representations

Outline of the Talk

- Background
  - Discriminative dependency parsing
  - Brown et al. (1992) algorithm
- Cluster-based feature sets
- Experiments on English and Czech

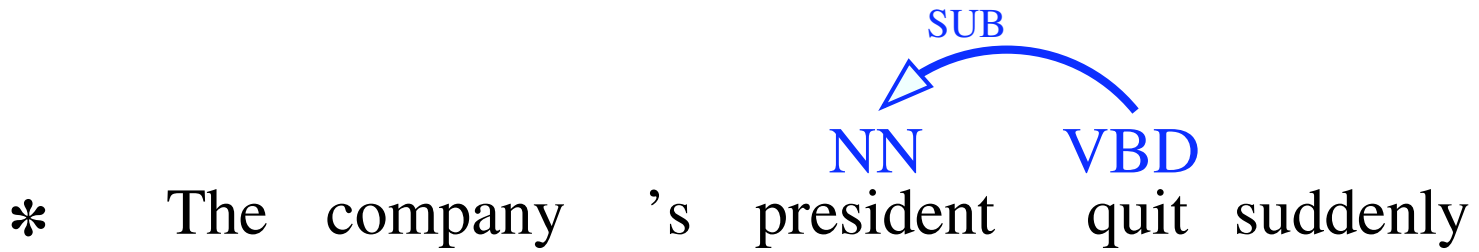
Review of Dependency Parsing



- Linear models for dependency parsing:

$$\textsc{Parse}(\mathbf{x}) = \underset{y \in \mathcal{Y}(\mathbf{x})}{\operatorname{argmax}} \ \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, y)$$

First-Order Models

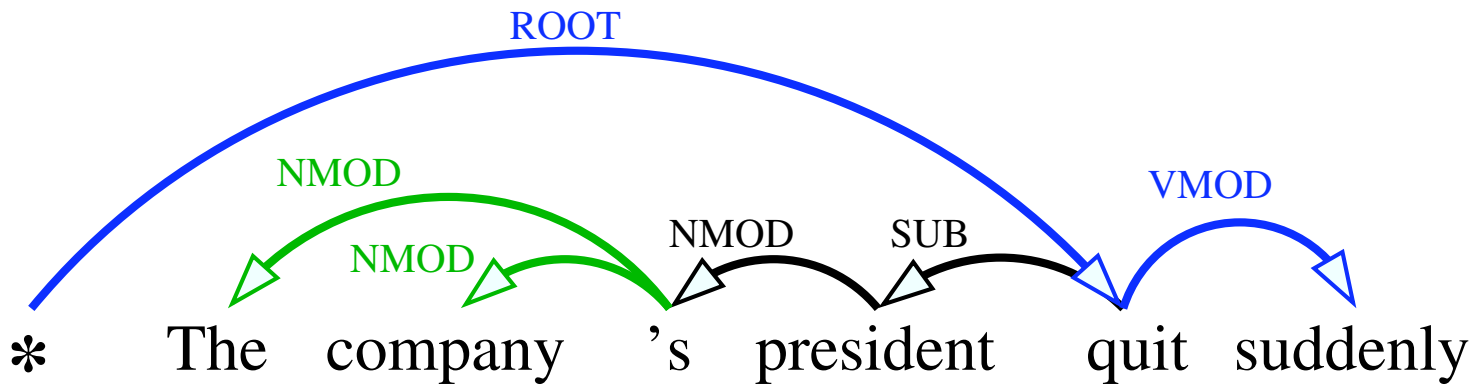


- First-order factorizations:

$$\textsc{Parse}(\mathbf{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}(\mathbf{x})} \sum_{d \in y} \mathbf{w} \cdot \mathbf{f}(\mathbf{x}, d)$$

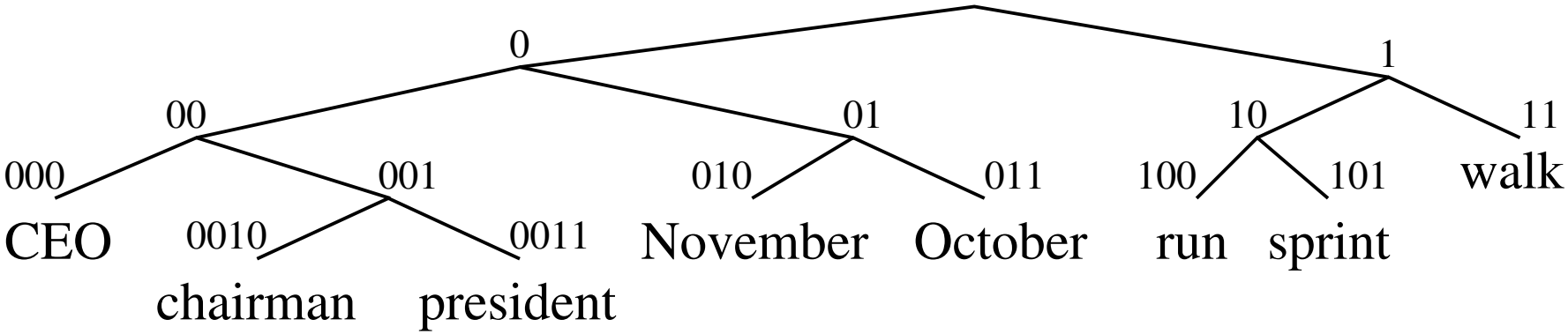
- Typical features are indicators for POS and word identity

Second-Order Models



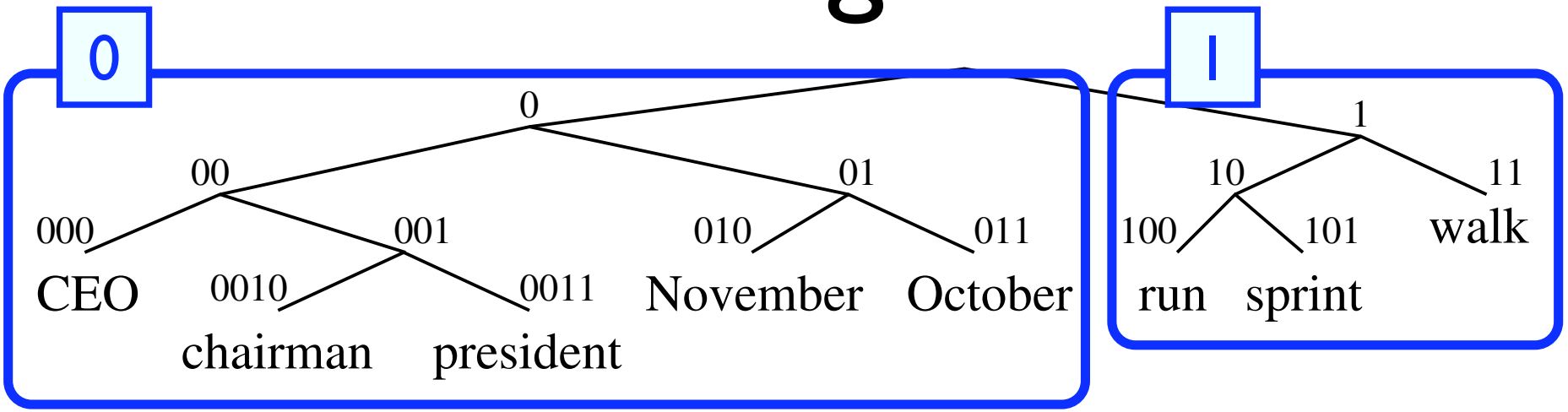
- Second-order models include sibling or grandparent dependencies

Brown Algorithm



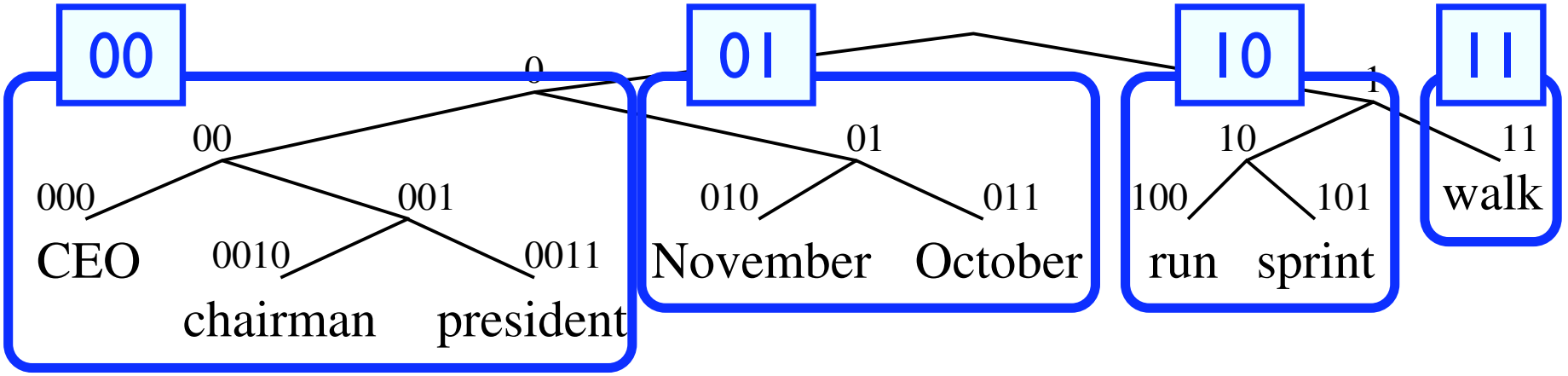
- Words merged according to contextual similarity
- Prefixes of bit strings yield clusterings
- Prefix length determines granularity

Brown Algorithm



- Words merged according to contextual similarity
- Prefixes of bit strings yield clusterings
- Prefix length determines granularity

Brown Algorithm



- Words merged according to contextual similarity
- Prefixes of bit strings yield clusterings
- Prefix length determines granularity

Brown Algorithm

- Examples of clusters from our English experiments

0101010101110011 constructed
0101010101110011 elucidated
0101010101110011 inhaled
0101010101110011 rewritten

Past Participle Verbs

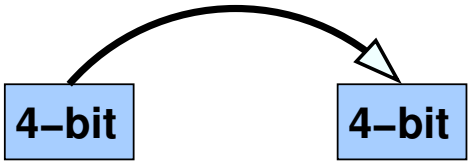
100111111111100 precious-metal
100111111111100 grain-futures
100111111111100 crude-oil-futures

Markets

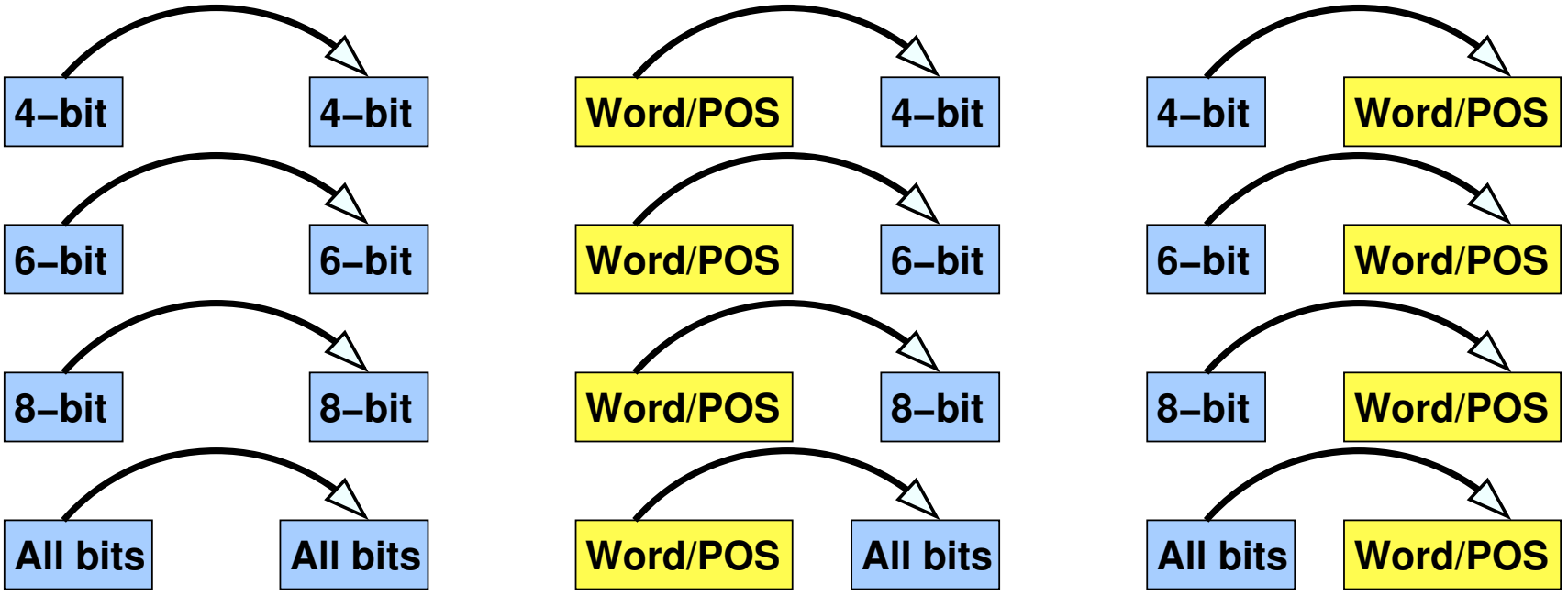
Cluster-based Features

- Discriminative parsers use feature vector representations, e.g., $\mathbf{f}(\mathbf{x}, d)$
- "Baseline" features include words and POS
- "Cluster-based" feature sets have additional templates with clusters

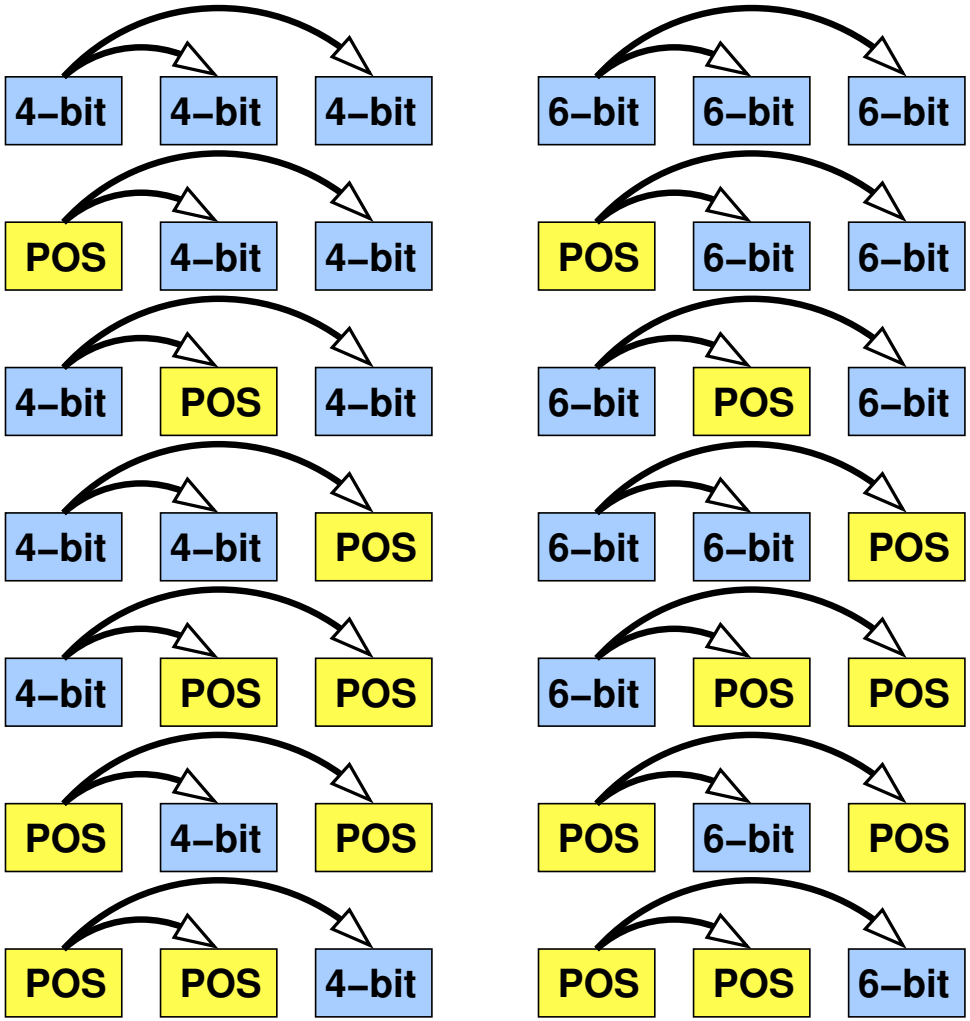
Cluster-based Features



Cluster-based Features



Cluster-based Features



Feature Pruning

- Cluster-based feature sets were very large
- Eliminate features using word frequency
  - Only use the top-800 most frequent words
  - Cluster-based features were not affected

Experiments

- Clusters obtained from implementation of Liang (2005)
- Averaged perceptron training
- First-order and second-order parsing
- Labeled and unlabeled parsing
- Baseline and cluster-based feature sets

English Experiments

- Penn Treebank
  - Train on Sections 2-21
  - Validate on Section 22
  - Test on Sections 0,1,23,24
- Clusters were derived from the BLLIP corpus (~43 million words)

English Baseline Model

Parsing Model	Accuracy
McDonald et al. (2006)	91.5
Second-order, baseline features	92.0

- Unlabeled head-prediction accuracy on Section 23
- Baseline model is state-of-the-art

English Labeled Parsing

Test Set	First-order parsing		Second-order parsing	
	Baseline	Cluster-based	Baseline	Cluster-based
Sec 00	90.3	91.0 (+0.7)	91.3	92.1 (+0.8)
Sec 01	90.8	91.7 (+0.9)	91.9	92.7 (+0.8)
Sec 23	90.3	91.2 (+0.9)	91.4	92.1 (+0.7)
Sec 24	89.6	90.1 (+0.5)	90.4	91.2 (+0.8)

- Labeled head-prediction accuracy on all test sets
- Cluster-based features outperform baseline

Effect of Training Corpus Size

- Examine the effect of the cluster-based features as the amount of training data varies
- High-quality POS tags are critical for parsing performance
- Two experimental settings

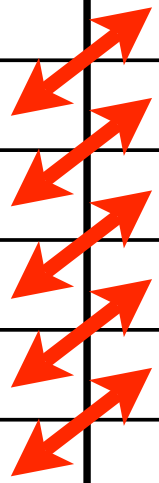
Effect of Training Corpus Size

Training Sentences	Baseline	Cluster-based
1000	86.3	87.5 (+1.2)
2000	87.7	88.9 (+1.2)
4000	89.2	90.5 (+1.3)
8000	90.6	91.6 (+1.0)
16000	91.3	92.4 (+1.1)
32000	92.1	93.4 (+1.3)
39832	92.4	93.3 (+0.9)

- Unlabeled accuracy on development set
- Tagger always trained on all 39832 sentences

Effect of Training Corpus Size

Training Sentences	Baseline	Cluster-based
1000	86.3	87.5 (+1.2)
2000	87.7	88.9 (+1.2)
4000	89.2	90.5 (+1.3)
8000	90.6	91.6 (+1.0)
16000	91.3	92.4 (+1.1)
32000	92.1	93.4 (+1.3)
39832	92.4	93.3 (+0.9)



- Unlabeled accuracy on development set
- Tagger always trained on all 39832 sentences

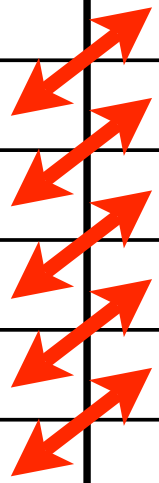
Effect of Training Corpus Size

Training Sentences	Baseline	Cluster-based
1000	82.0	85.3 (+3.3)
2000	85.0	87.5 (+2.5)
4000	87.9	89.7 (+1.8)
8000	89.7	91.4 (+1.7)
16000	91.1	92.2 (+1.1)
32000	92.1	93.2 (+1.1)
39832	92.4	93.3 (+0.9)

- Unlabeled accuracy on development set
- Tagger and parser trained on small datasets

Effect of Training Corpus Size

Training Sentences	Baseline	Cluster-based
1000	82.0	85.3 (+3.3)
2000	85.0	87.5 (+2.5)
4000	87.9	89.7 (+1.8)
8000	89.7	91.4 (+1.7)
16000	91.1	92.2 (+1.1)
32000	92.1	93.2 (+1.1)
39832	92.4	93.3 (+0.9)



- Unlabeled accuracy on development set
- Tagger and parser trained on small datasets

Czech Experiments

- Prague Dependency Treebank
  - Train/val/test as provided in the corpus
- Clusters were derived from the unlabeled text portion of the PDT (~39 million words)
- First-order non-projective parsing (MST), second-order projective parsing

Czech Unlabeled Parsing

Parsing Model	Baseline	Cluster-based
First-order MST	84.5	86.1 (+1.6)
Second-order	86.1	87.1 (+1.0)

Related Work	Accuracy
McDonald (2005) first-order MST	84.4
McDonald (2006) second-order	85.2
Nivre and Nilsson (2005)	80.1
Hall and Novák (2005)	85.1

- Unlabeled head-prediction accuracy on test set

Effect of Training Corpus Size

Training Sentences	Baseline	Cluster-based
1000	74.4	74.6 (+0.2)
2000	76.6	77.6 (+1.0)
4000	78.3	79.3 (+1.0)
8000	79.8	81.0 (+1.2)
16000	82.5	83.7 (+1.2)
32000	84.7	85.8 (+1.1)
64000	86.0	87.1 (+1.1)
73088	86.1	87.3 (+1.2)

- Unlabeled accuracy on development set

Feature Pruning

Word Threshold	Baseline	Cluster-based
100	90.6	93.1
200	91.4	93.2
400	91.7	93.2
800	91.9	93.3
1600	92.2	
All words	92.4	

- Unlabeled accuracy on development set
- Cluster-based features are stable

Effect of POS Tags

Features	First-order	Second-order
no POS, no clusters	77.2	86.7
no POS, with clusters	90.7	91.8
with POS, no clusters	90.9	92.4

- Unlabeled accuracy on development set
- Clusters *alone* are almost as good as baseline

Conclusions

- Lexical statistics are important but sparse
- Word clusters serve as an alternate lexical representation
- Clusters incorporated as features for a discriminative parser
- Performance gains over a state-of-the-art baseline model