JOHNS HOPKINS UNIVERSITY

Microsoft Research

Low-Resource Semantic Role Labeling

Matthew R. Gormley
Margaret Mitchell
Benjamin Van Durme
Mark Dredze

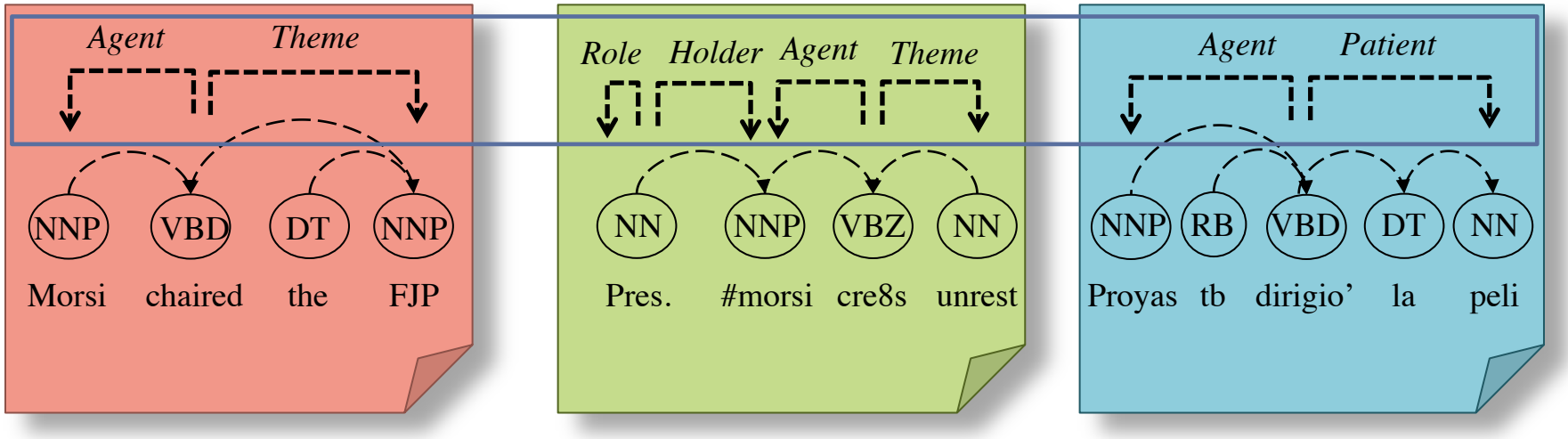
See our paper from ACL '14

CLSP Seminar ; September 12, 2014

Shallow Semantics

Our End Task:

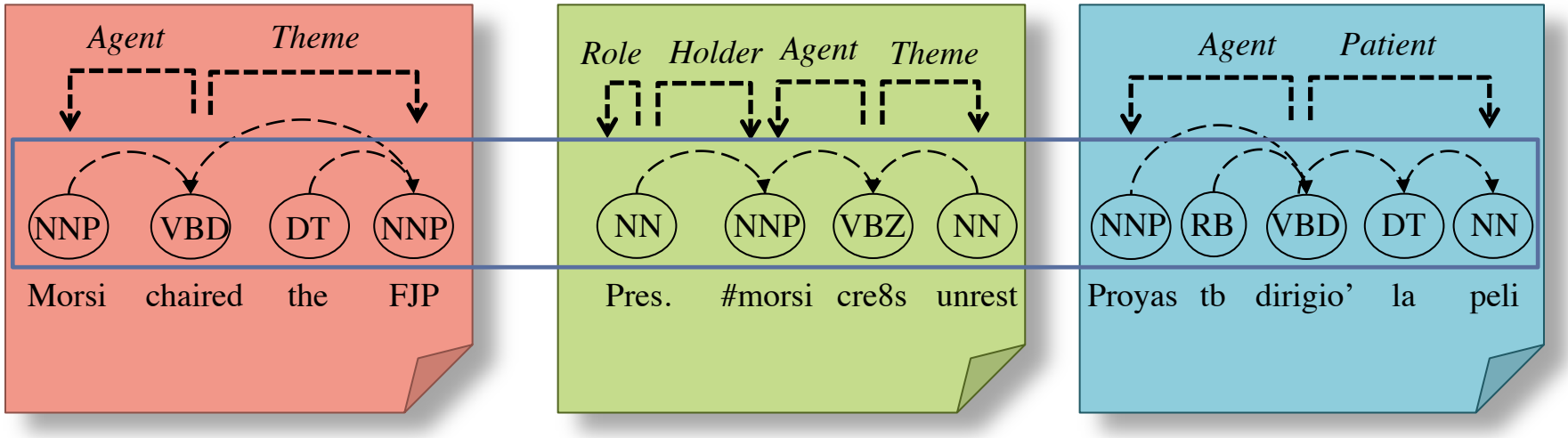
- Representation: **Semantic Role Labeling** (SRL)
  - Intuitively captures *who* did *what* to *whom*, *when* and *where*
  - Similar to open-domain relation extraction
- Languages: Catalan, Czech, German, English, Spanish, Chinese



Syntax

Intermediate Tasks:

- Dependency parsing
  - Captures the structure of the sentence
  - Diverges from the shallow semantics representation
- Part-of-speech (POS) tagging

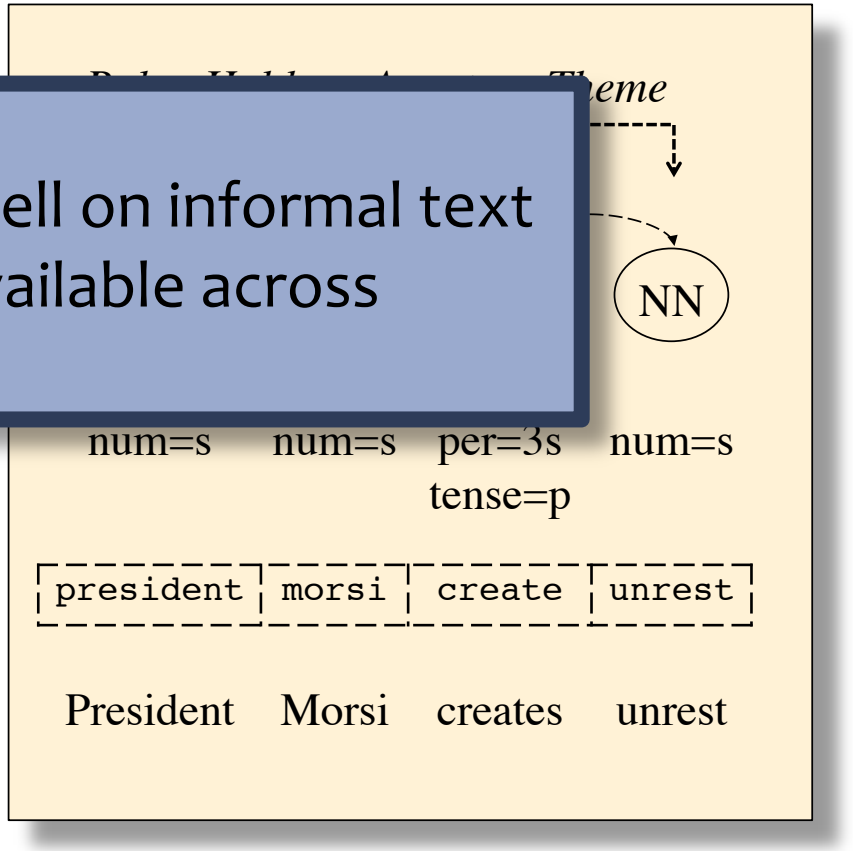


Background: The Supervised SRL Pipeline

Pipelined Training: Train each component of the pipeline independently using the predictions of the previous stage(s) as features.

Tool
Semantic ro
Dependenc
Part-of-spee
Morphological feature extractor
Lemmatizer

- Costly
- Does not work well on informal text
- Resources not available across languages

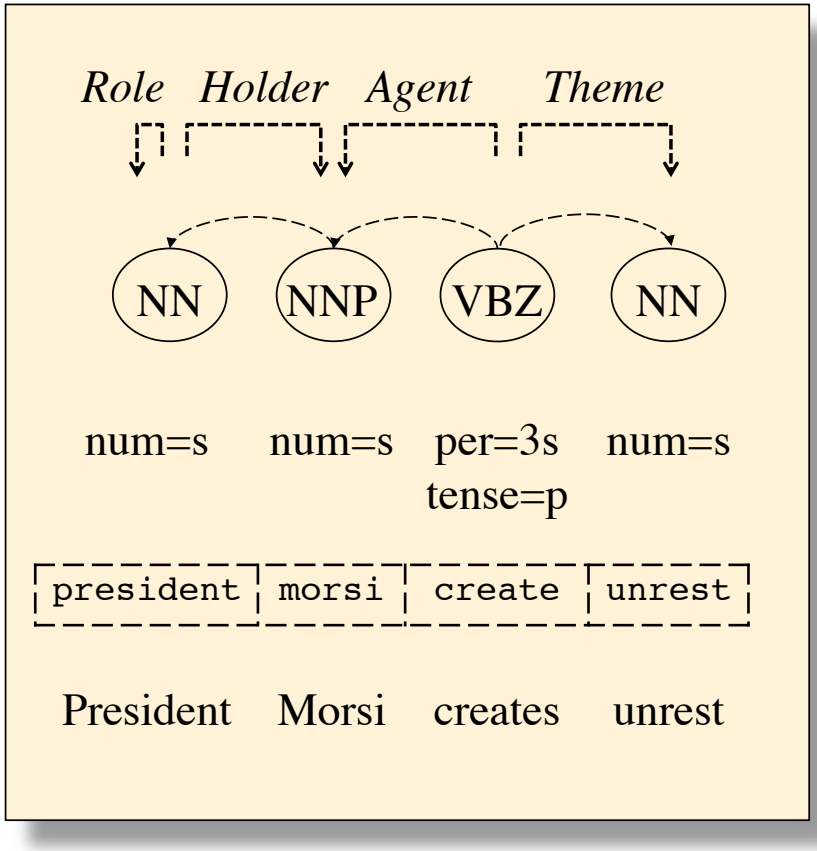


Background: The Supervised SRL Pipeline

Pipelined Training: Train each component of the pipeline independently using the predictions of the previous stage(s) as features.

Our Emphasis

Tool
Semantic role labeler
Dependency parser
Part-of-speech tagger
Morphological feature extractor
Lemmatizer



This Talk in a Nutshell

- We want to do SRL in a new language.
- Syntax helps, but is very expensive to annotate.

Supervised Annotation	Cost
Semantic roles	$$$
Dependency parses	$$$$$$$
Part-of-speech (POS) tags	$$
Morphology	$$$
Lemmas	$

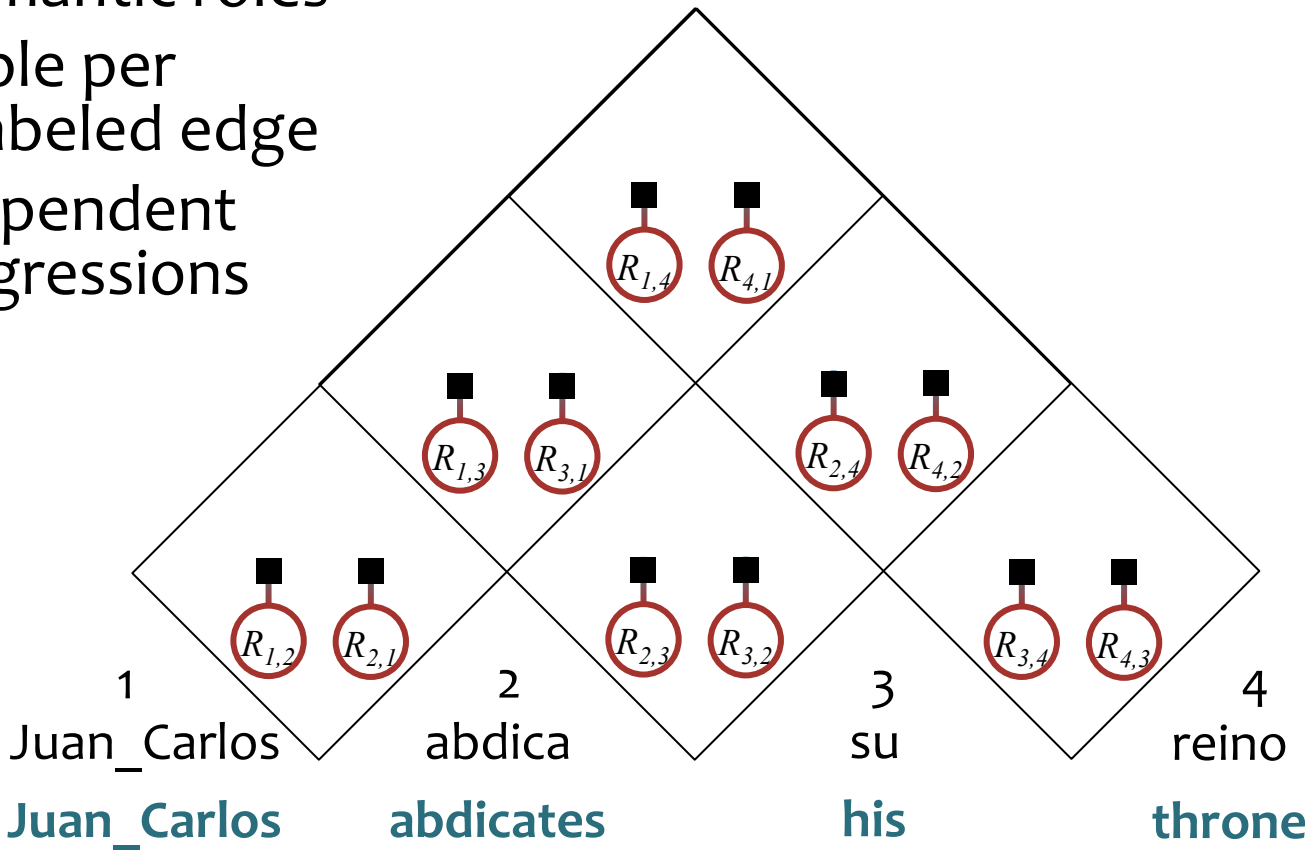
This Talk in a Nutshell

- We want to do SRL in a new language.
- Syntax helps, but is very expensive to annotate.
- Having annotated syntax would be nice, but **we can make progress without it**!

Supervised Annotation	Cost
Semantic roles	$$$
~~Dependency parses~~	~~$$$$$$$~~
~~Part of speech (POS) tags~~	~~$$~~
~~Morphology~~	~~$$$~~
~~Lemmas~~	~~$~~

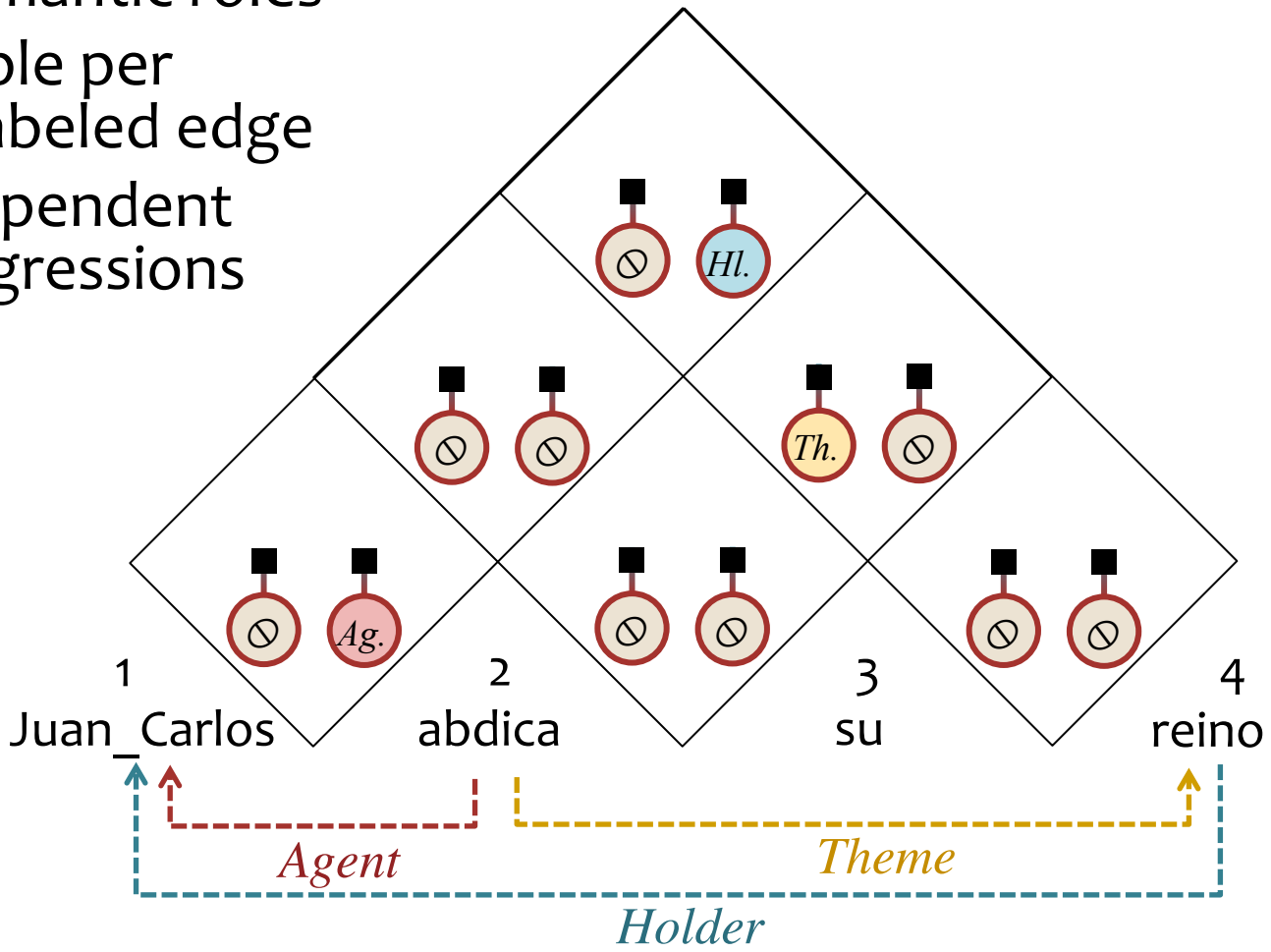
Supervised SRL as a Factor Graph

- Jointly *identify* and *classify* semantic roles
- One variable per possible labeled edge
- $O(n^2)$ independent logistic regressions



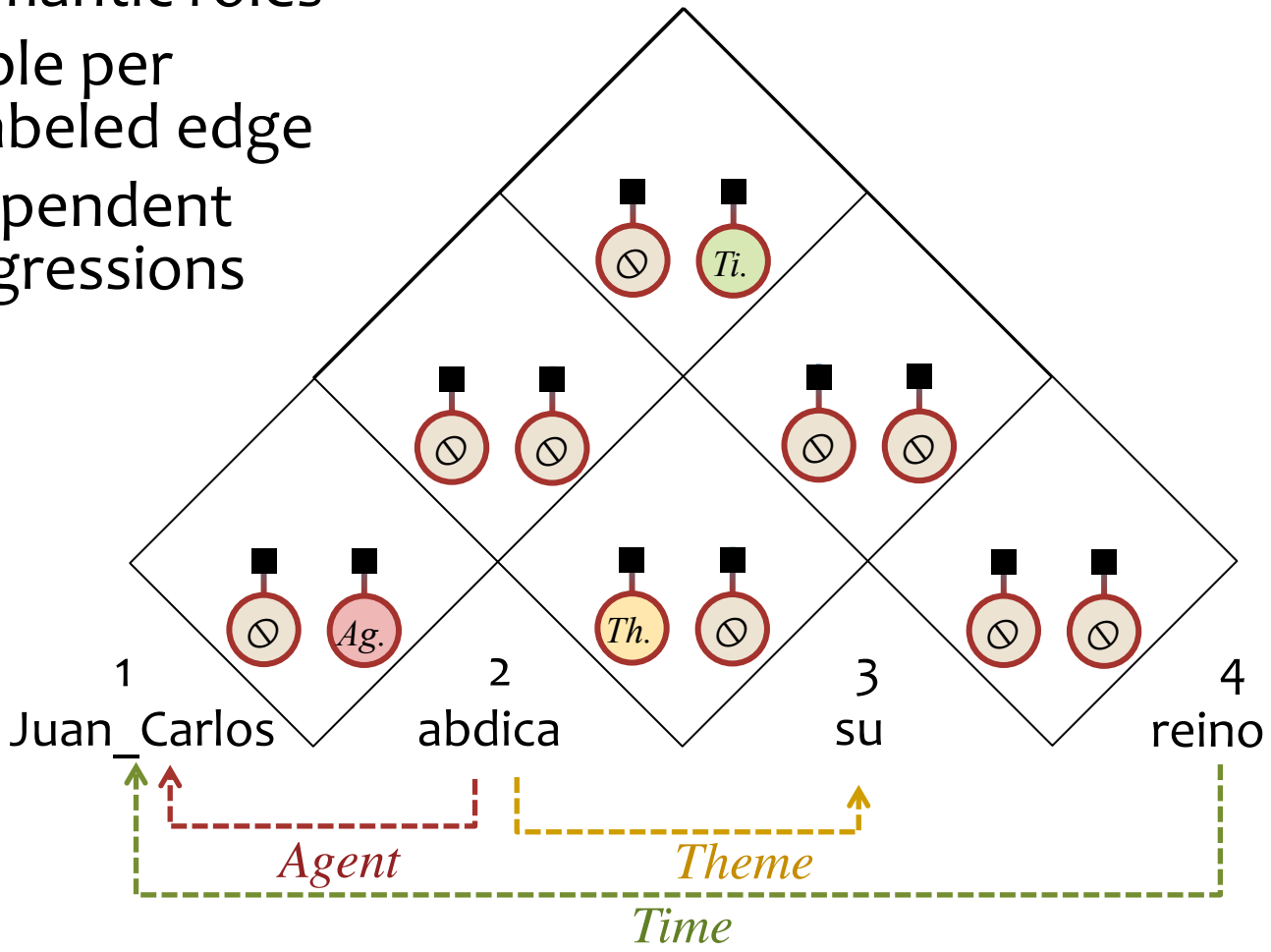
Supervised SRL as a Factor Graph

- Jointly *identify* and *classify* semantic roles
- One variable per possible labeled edge
- $O(n^2)$ independent logistic regressions



Supervised SRL as a Factor Graph

- Jointly *identify* and *classify* semantic roles
- One variable per possible labeled edge
- $O(n^2)$ independent logistic regressions

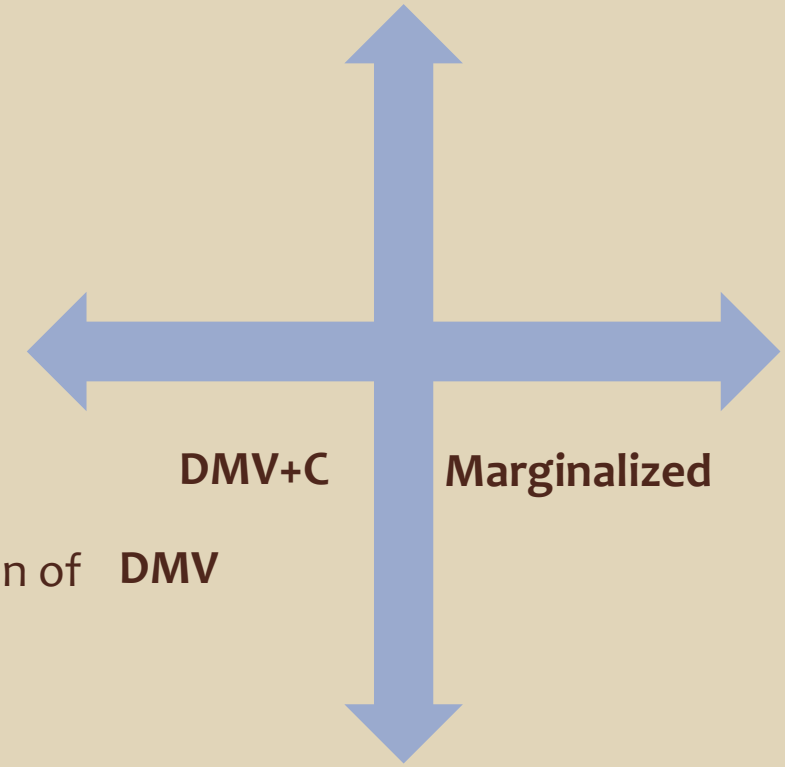


High Resource

Pro: high accuracy parsers
Con: expensive

Pipeline

Pro: easy to throw in lots of features
Con: propagation of errors



Joint

Pro: confidence flows between levels of the model
Con: features must permit efficient inference

Low Resource

Pro: cheap, easily deployable
Con: low accuracy latent syntax

Contributions

- **Experimental contributions:**
  - Comparison of **pipeline** and **joint** models for SRL.
  - **Subtractive experiments** that consider the removal of supervised data.
  - Analysis of the induced grammars in (1) unsupervised, (2) distantly-supervised, and (3) joint training settings.
- **Modeling contributions:**
  - **Simpler joint CRF** for syntactic and semantic dependency parsing than previously reported.
  - **New application** of unsupervised **grammar induction**: low-resource SRL.
  - **Constrained grammar induction** using SRL for distant-supervision.
  - Use of **Brown clusters** in place of POS tags for low-resource SRL.

Three Training Settings for Latent Syntax

1. Fully Unsupervised (DMV)
2. Distantly Supervised (DMV+C)
3. Jointly Learned with SRL (Marginalized)

Latent Syntax

1. Fully Unsupervised

A. Brown clusters (Brown et al., 1992) in place of **POS** tags

- Clusters formed by hierarchical clustering; maximizes likelihood under latent-class bigram model

B. Syntax from **D**ependency **M**odel with **V**alence (DMV) (Klein & Manning, 2004)

- Children generated recursively
- Viterbi EM training (Spitkovsky et al., 2010)

Latent Syntax

2. Distantly Supervised

- Viterbi EM training of DMV
- Observes semantic graph during training
- Constrain a CKY parser in E-step to respect SRL

Algorithm:

1. Define DMV as a PCFG (Cohn et al., 2010)
2. CKY parse (Younger, 1967; Aho and Ullman, 1972)
3. Populate cells with SRL-compatible non-terminals

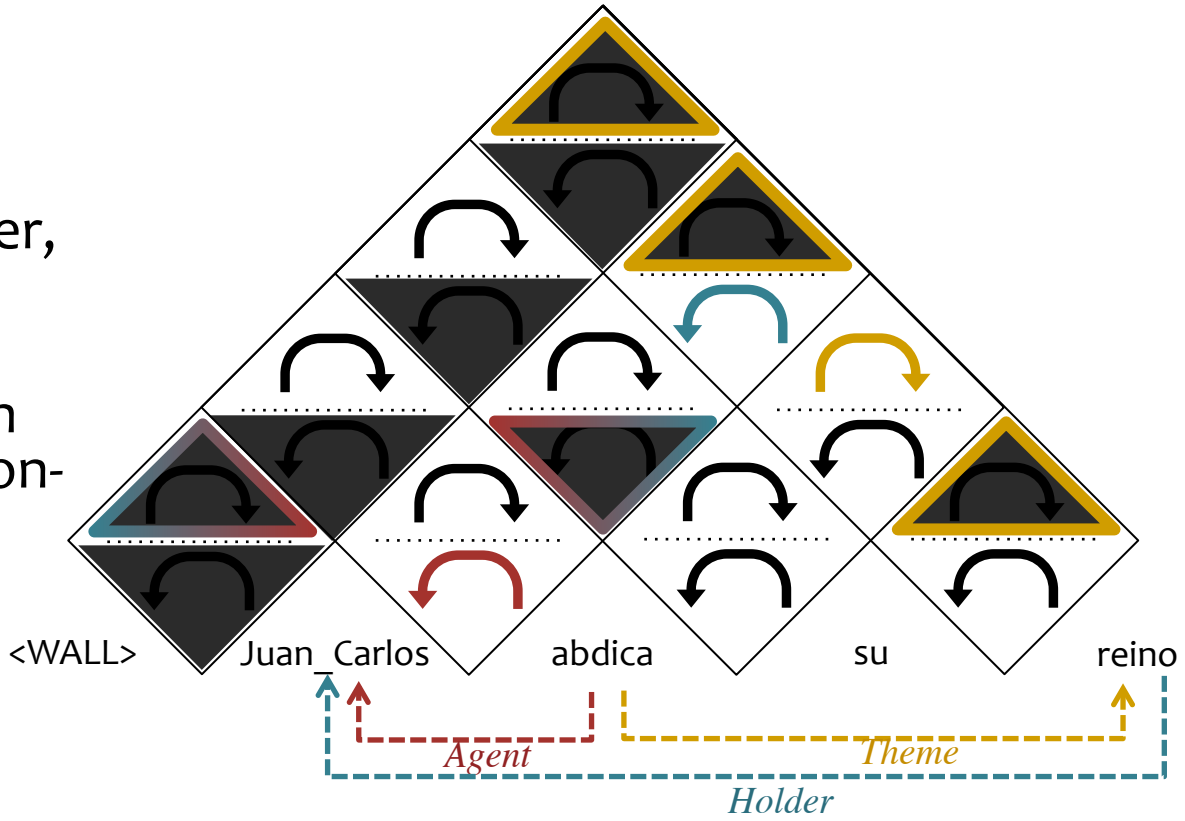
Latent Syntax

2. Distantly Supervised

- Viterbi EM training of DMV
- Observes semantic graph during training
- Constrain a CKY parser in E-step to respect SRL

Algorithm:

1. Define DMV as a PCFG (Cohn et al., 2010)
2. CKY parse (Younger, 1967; Aho and Ullman, 1972)
3. Populate cells with SRL-compatible non-terminals



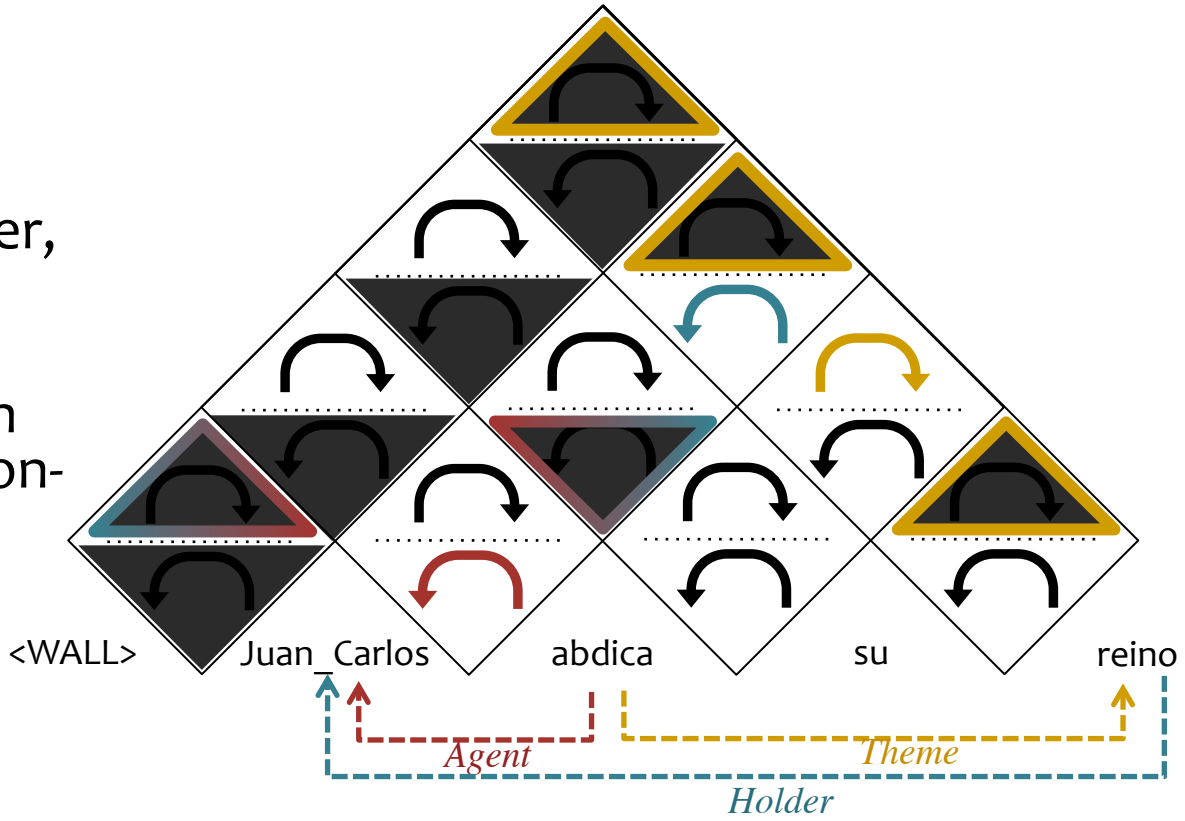
Latent Syntax

2. Distantly Supervised

- Viterbi EM training of DMV
- Observes semantic graph during training
- Constrain a CKY parser in E-step to respect SRL

Algorithm:

1. Define DMV as a PCFG (Cohn et al., 2010)
2. CKY parse (Younger, 1967; Aho and Ullman, 1972)
3. Populate cells with SRL-compatible non-terminals



Latent Syntax

3. Joint Model

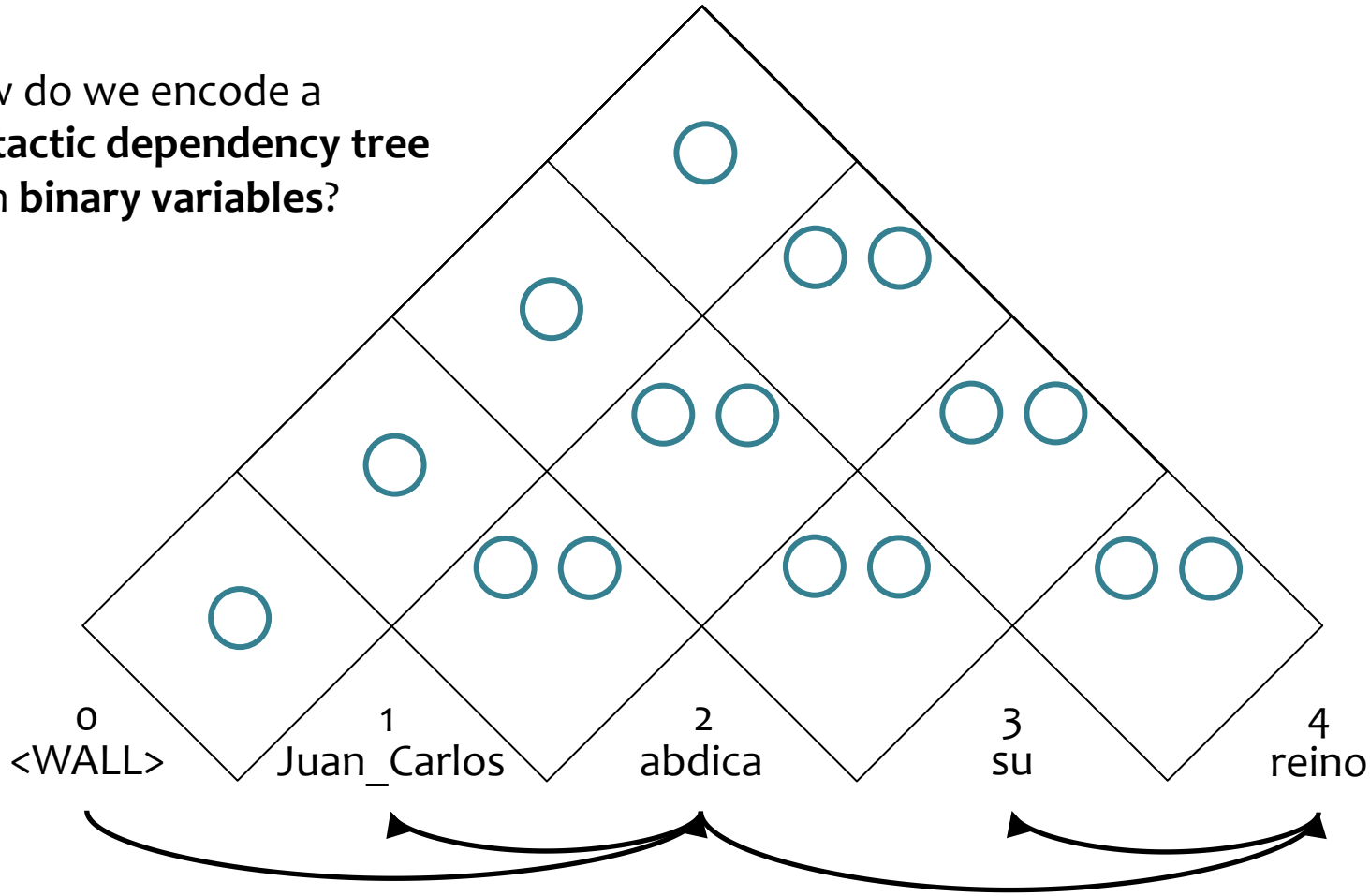
Marginalize over latent syntax to find the optimal semantic role assignment

- **Model:** Slight simplification of Naradowsky et al. (2012). Jointly *identify* and *classify* semantic roles.
- **Inference:** Belief propagation with inside-outside algorithm embedded in global factor (Smith & Eisner, 2008)
- **Brown clusters** in place of **POS** tags

Latent Syntax

3. Joint Model

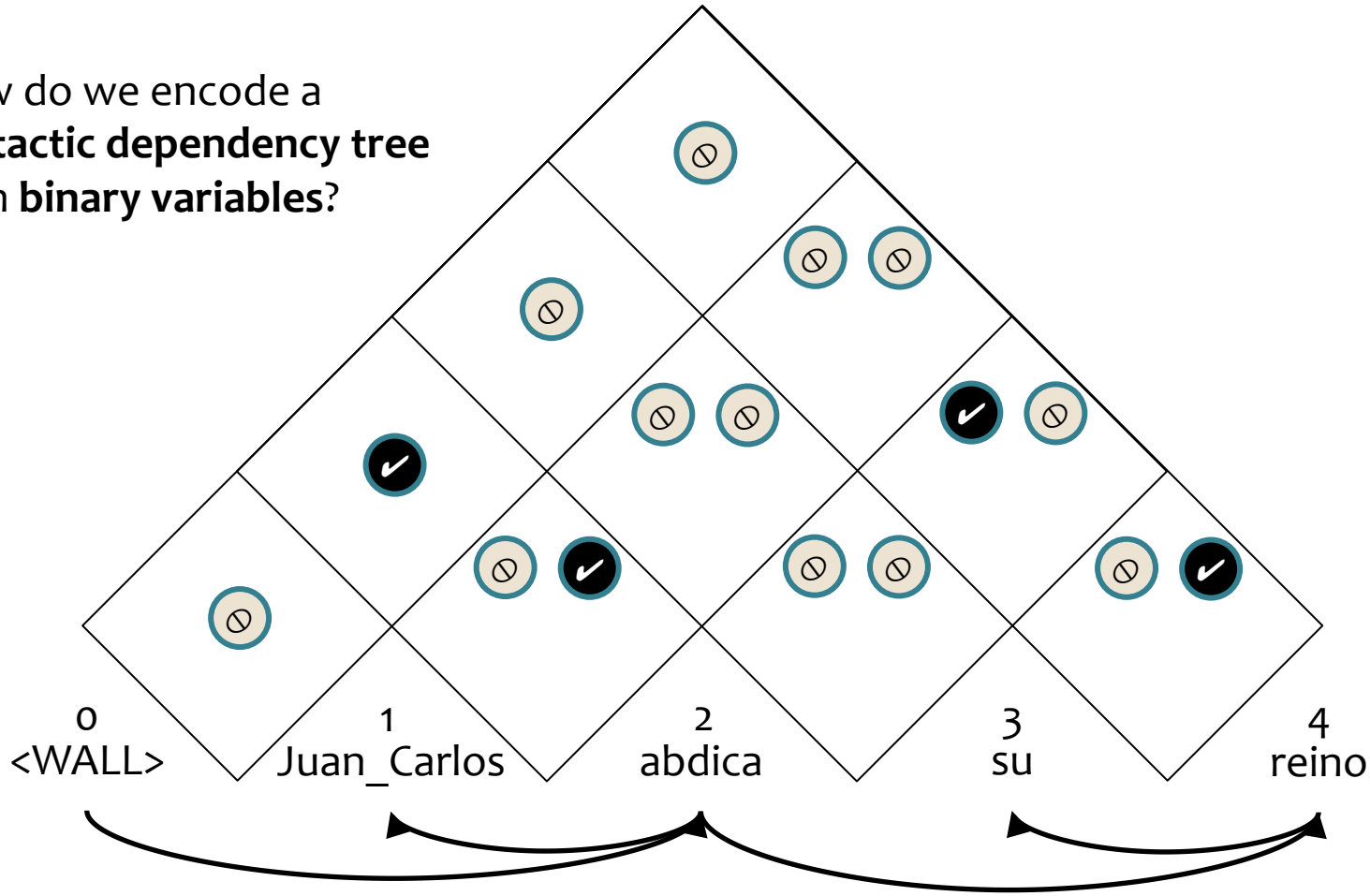
How do we encode a **syntactic dependency tree** with **binary variables**?



Latent Syntax

3. Joint Model

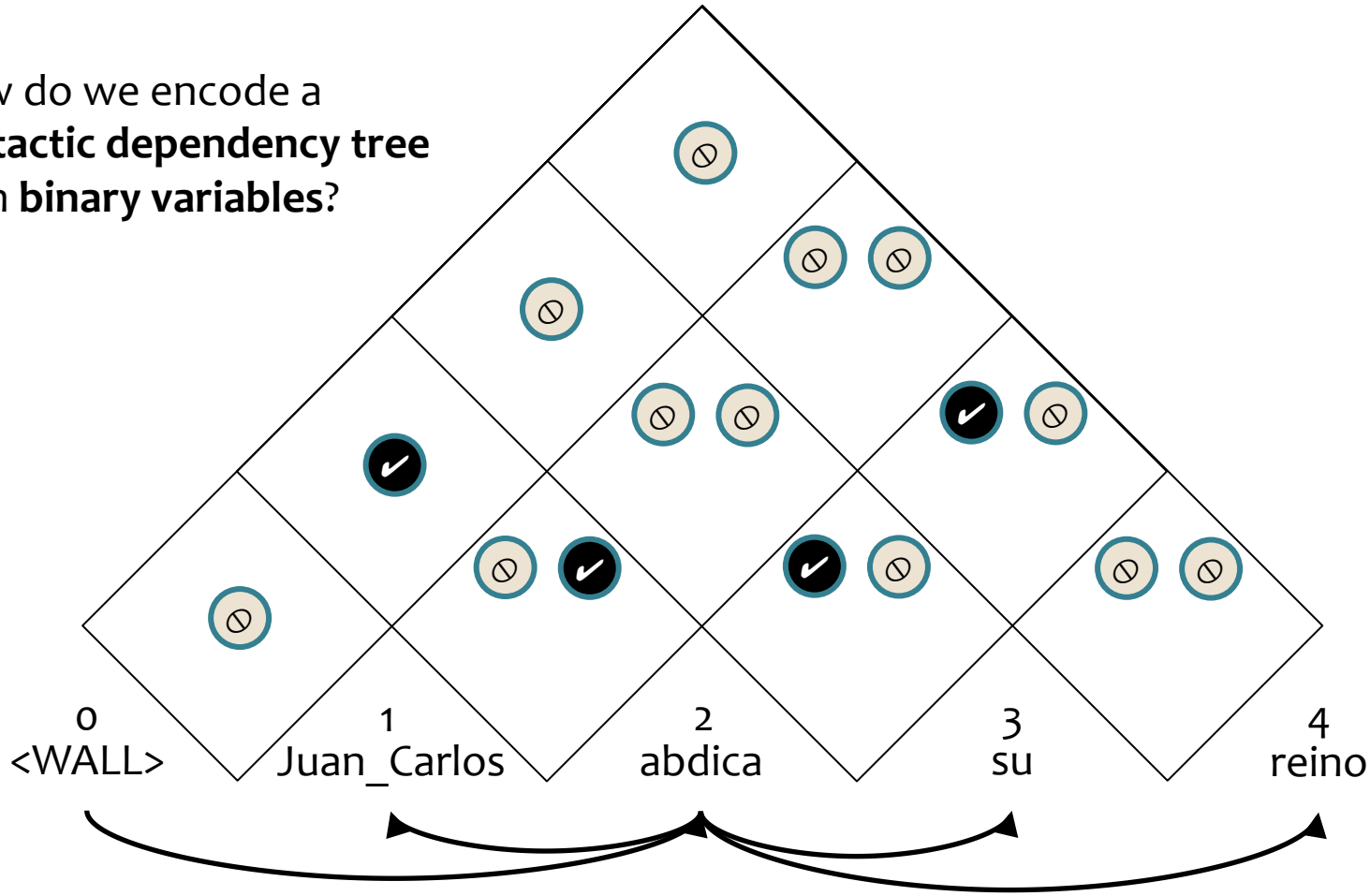
How do we encode a **syntactic dependency tree** with **binary variables**?



Latent Syntax

3. Joint Model

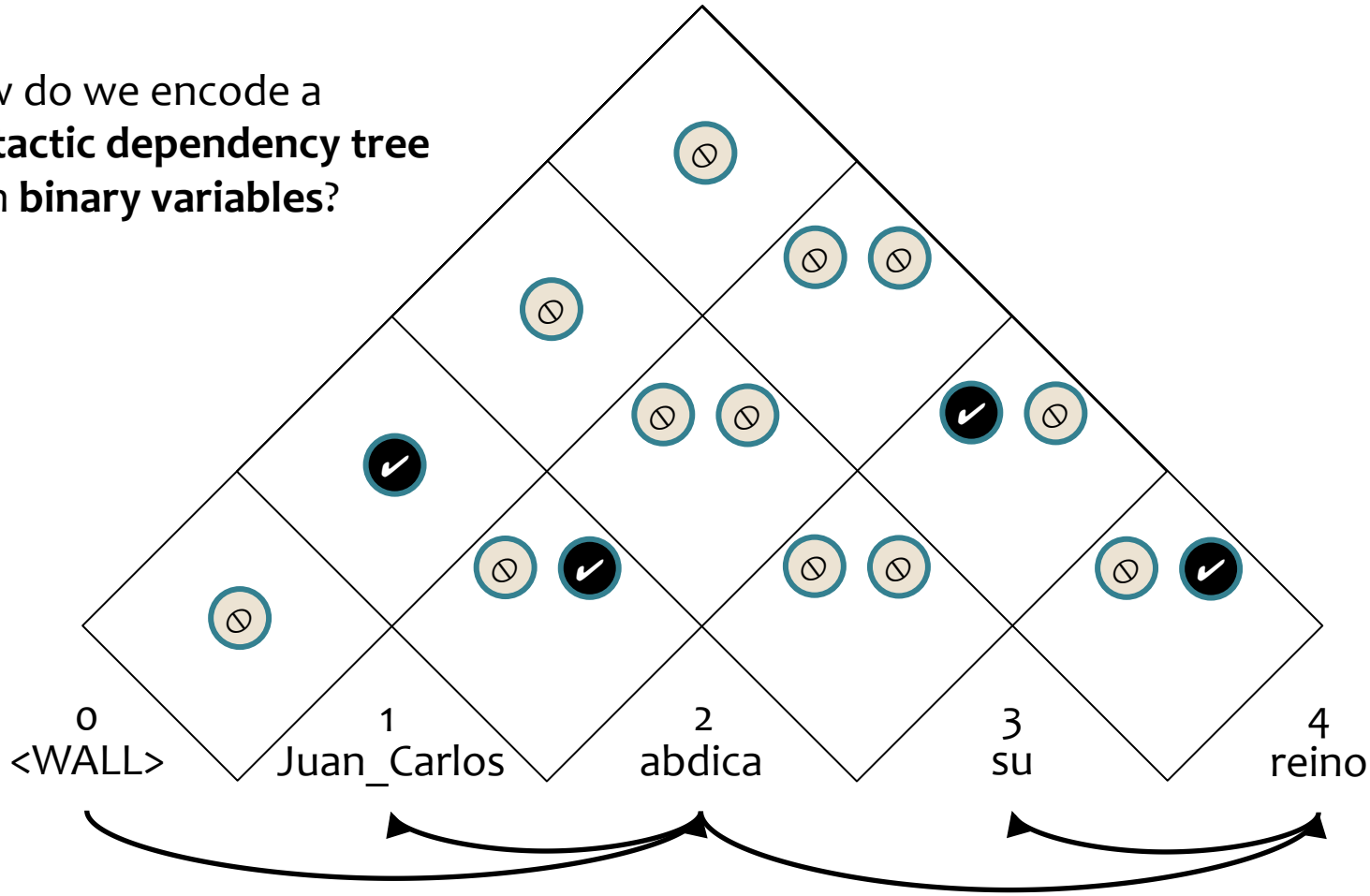
How do we encode a **syntactic dependency tree** with **binary variables**?



Latent Syntax

3. Joint Model

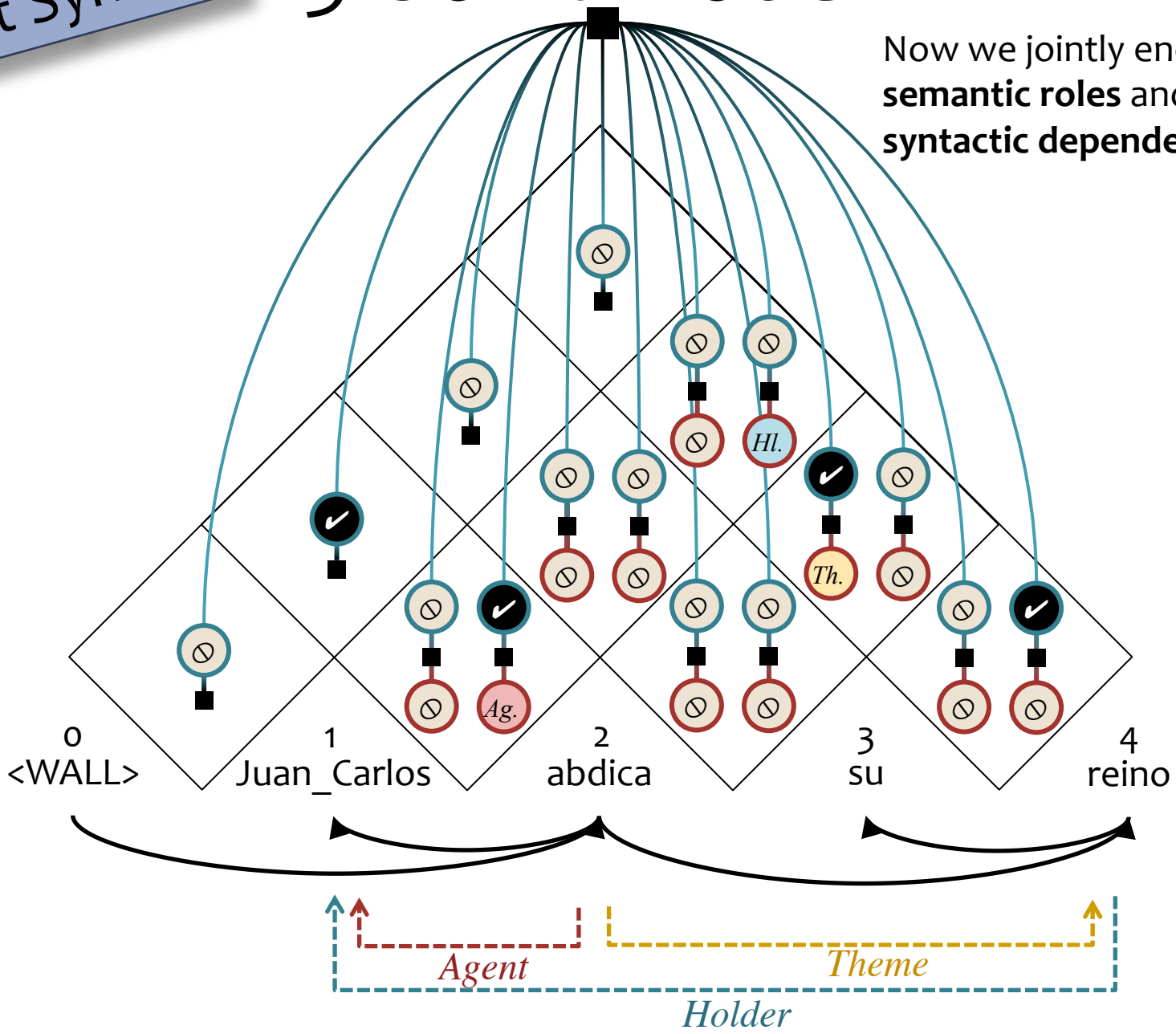
How do we encode a **syntactic dependency tree** with **binary variables**?



Latent Syntax

3. Joint Model

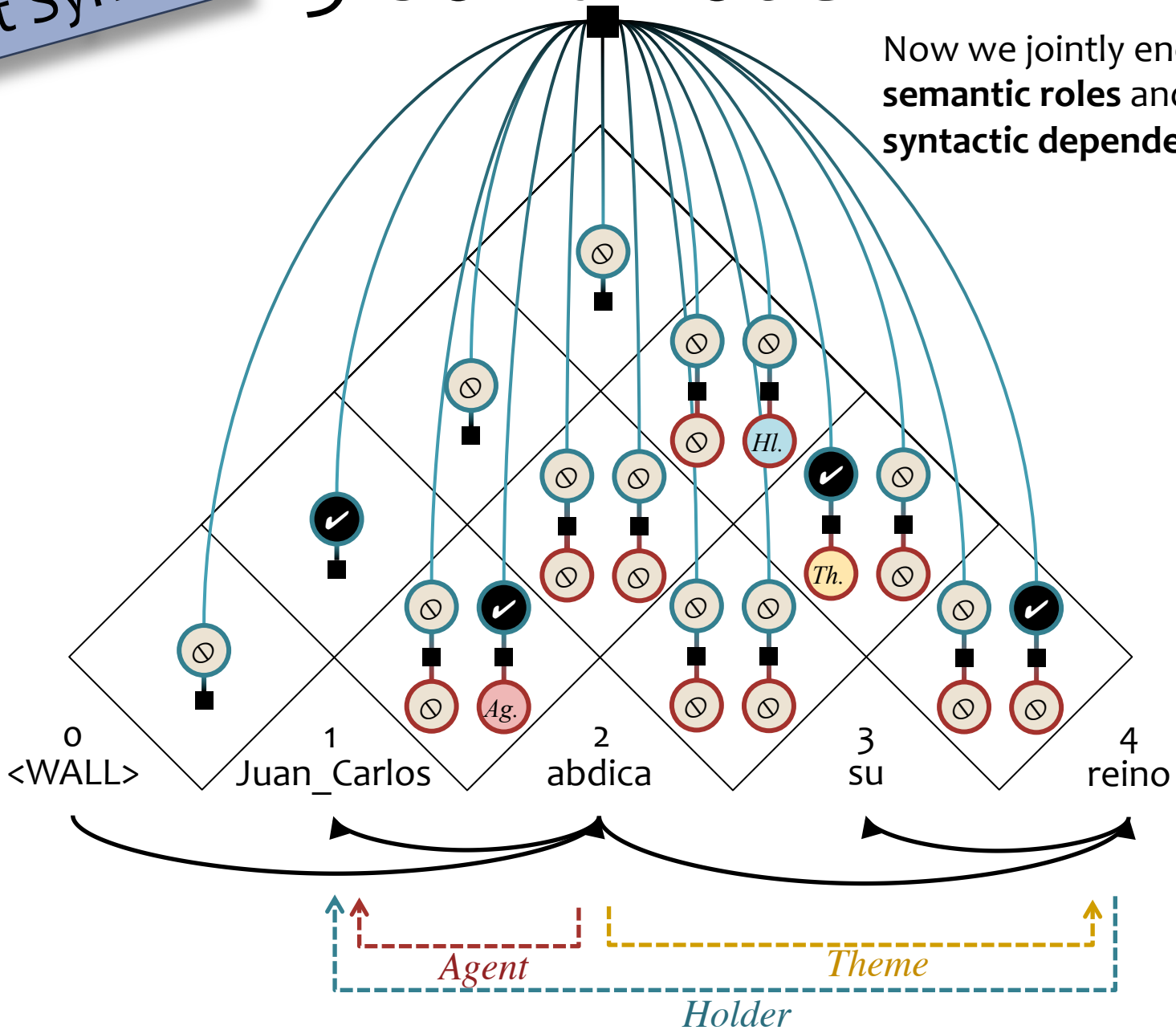
Now we jointly encode the **semantic roles** and the **syntactic dependency tree.**



Latent Syntax

3. Joint Model

Now we jointly encode the **semantic roles** and the **syntactic dependency tree.**



Features and Feature Selection

- Define millions of features using 100+ feature **templates**
- Incorporate feature ideas from:
  - Koo et al. (2008)
  - Björkelund et al. (2009)
  - Zhao et al. (2009)
  - Lluís et al (2013)

What about pairs of unigram templates?

Unigram Templates:

```
word(p)
lemma(p)
pos(p)
bc0(p)
bc1(p)
morpho(p)
deprel(p)
lc(p)
chpre5(p)
capitalized(p)
wordTopN(p)
morpho1(p)
morpho2(p)
morpho3(p)
eachmorpho(p)
```

Features and Feature Selection

Use Information Gain (IG) to find top unigram templates (Martins et al., 2011)

$$IG_{a,m} = \sum_{f \in T_m} \sum_{x_a} p(f, x_a) \log_2 \frac{p(f, x_a)}{p(f)p(x_a)}$$

Then combine top unigram templates to find top **bigram** templates.

Experiments

Datasets:

- Semantic Roles:
  - CoNLL-2009 Shared Task
  - Languages: Catalan, Czech, German, English, Spanish, Chinese
- Grammar Induction:
  - Additional experiments on WSJ portion of Penn Treebank for comparability
- Brown Clusters:
  - Wikipedia

CoNLL-2009 Supervised Data
Semantic roles
Dependency parses
Part-of-speech tags
Morphological features
Lemmas

Experiments

- Abbreviations for Latent Syntax:
  1. Fully Unsupervised (DMV)
  2. Distantly Supervised (DMV+C)
  3. Jointly Learned with SRL (Marginalized)
- Abbreviations for Tag Types:
  1. Part-of-speech Tags (pos)
  2. Brown Clusters (bc)

Grammar Induction Analysis

Does the latent syntax look any good?

Dependency Parser	Avg UAS	Catalan	Czech	German	English	Spanish	Chinese
Supervised	87.1	89.4	85.3	89.6	88.4	89.2	80.7

Grammar Induction Analysis

Does the latent syntax look any good?

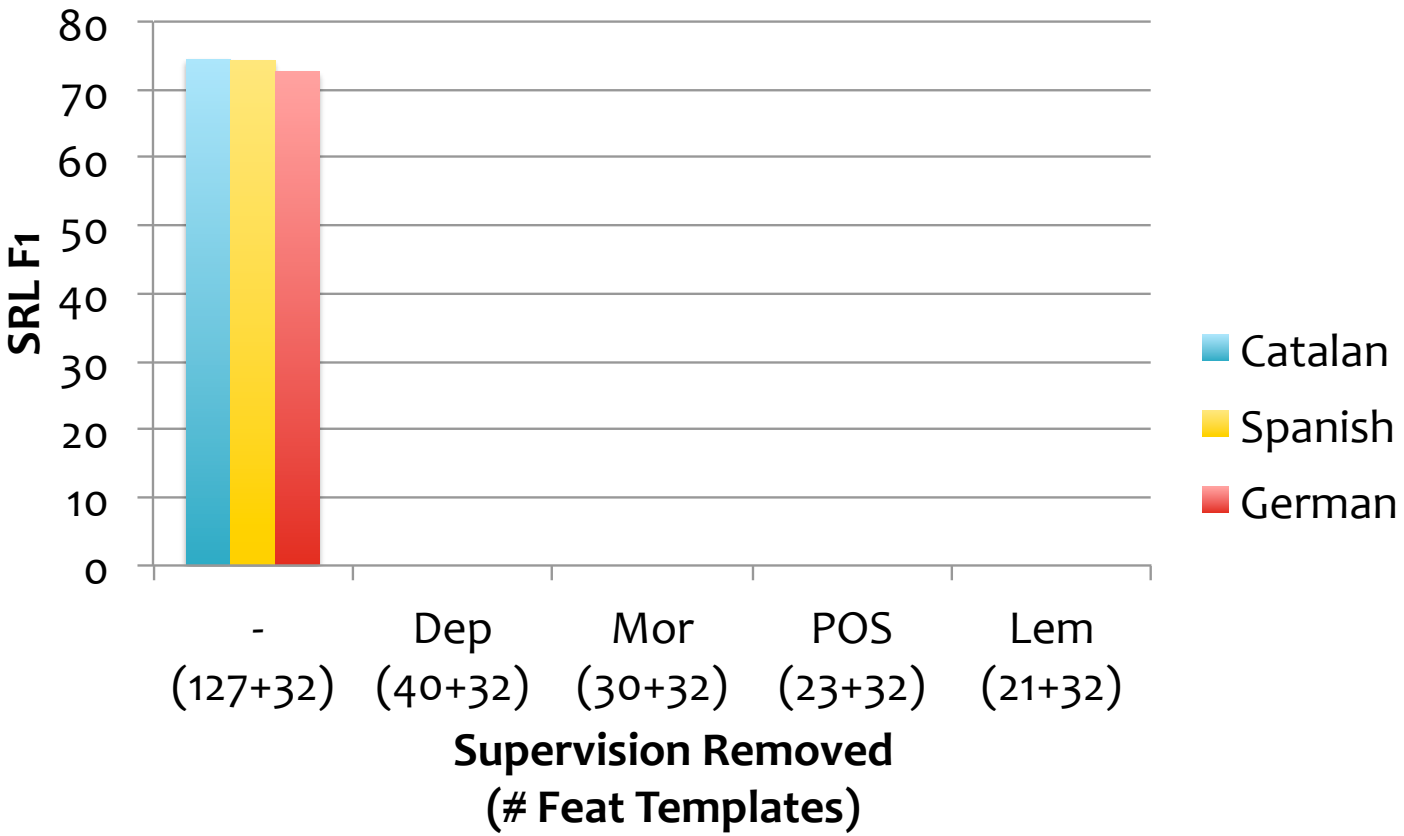
Dependency Parser	Avg UAS	Catalan	Czech	German	English	Spanish	Chinese
Supervised	87.1	89.4	85.3	89.6	88.4	89.2	80.7
Marginalized, IG_B	50.2	52.4	43.4	41.3	52.6	55.2	56.2
Marginalized, IG_c	43.8	50.3	45.8	27.2	44.2	46.3	48.5
DMV+C (bc)	40.2	46.3	37.5	28.7	40.6	50.4	37.5
DMV+C (pos)	37.5	50.2	34.9	21.5	36.9	49.8	32
DMV (pos)	30.2	45.3	22.7	20.9	32.9	41.9	17.2
DMV (bc)	22.1	18.8	32.8	19.6	22.4	20.5	18.6

Gap between supervised and *not-so-supervised* parser is very large

Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

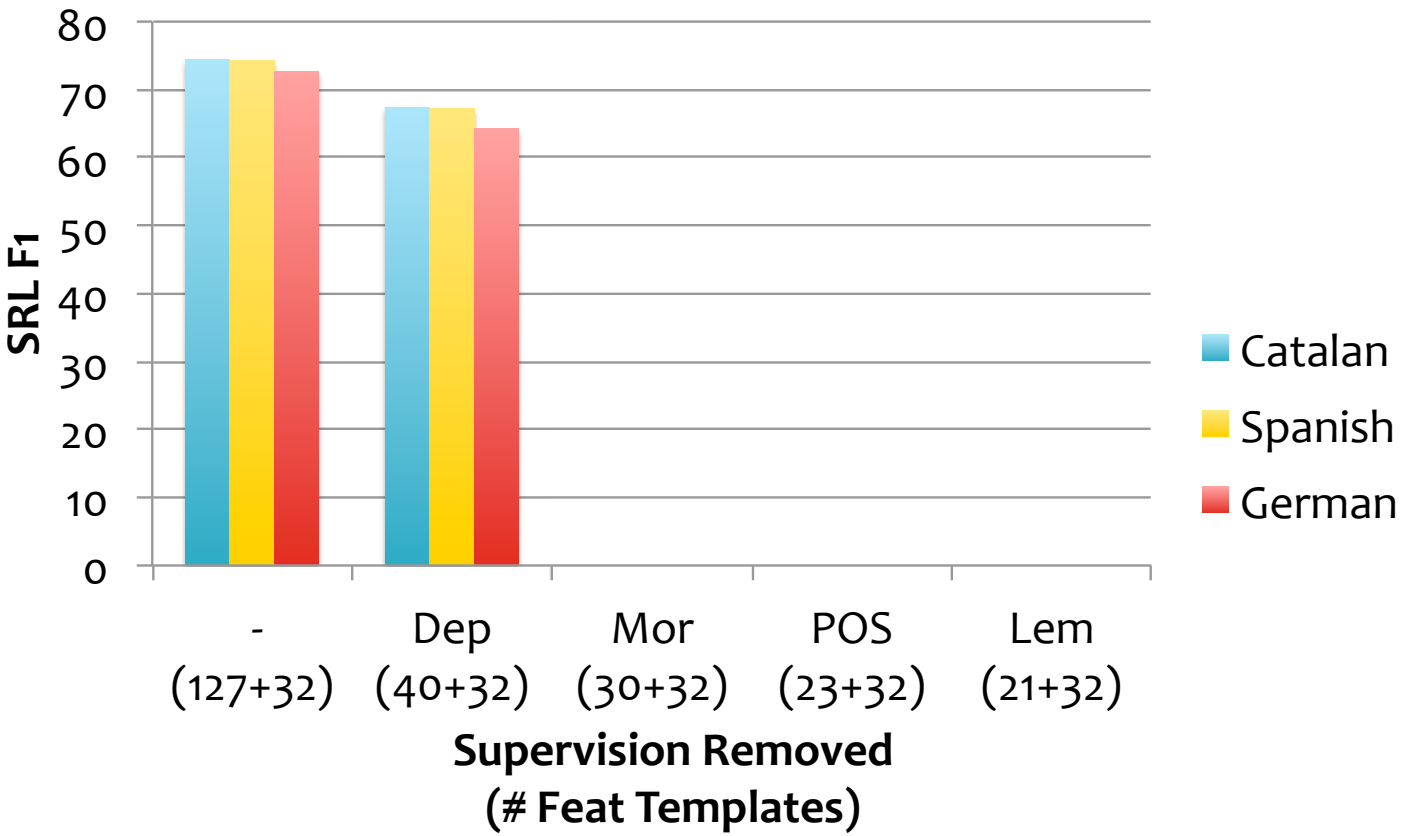
CoNLL-2009 Supervised Data
Semantic roles
Dependency parses
Morphology
Part-of-speech tags
Lemmas



Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

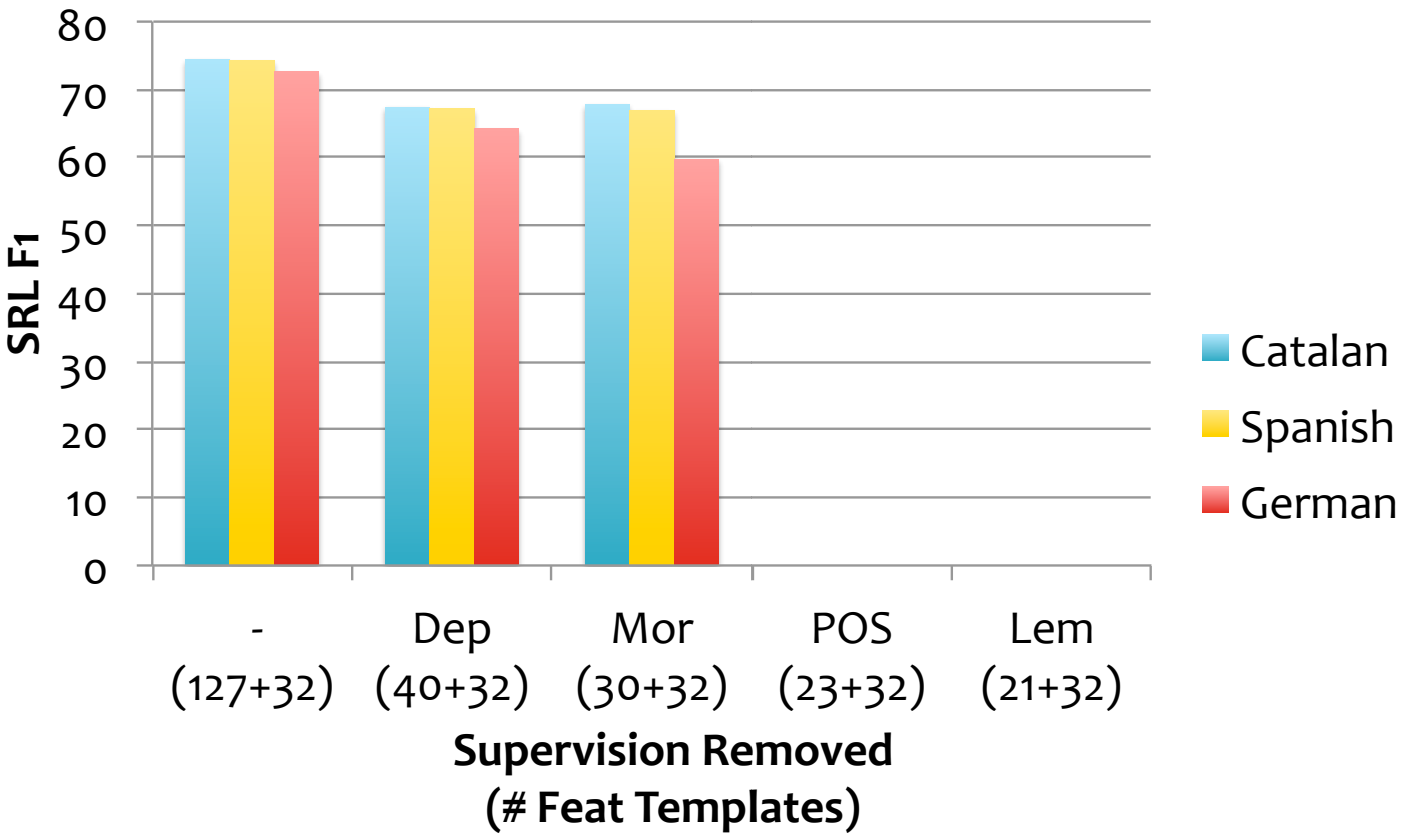
CoNLL-2009 Supervised Data
Semantic roles
~~Dependency parses~~
Morphology
Part-of-speech tags
Lemmas



Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

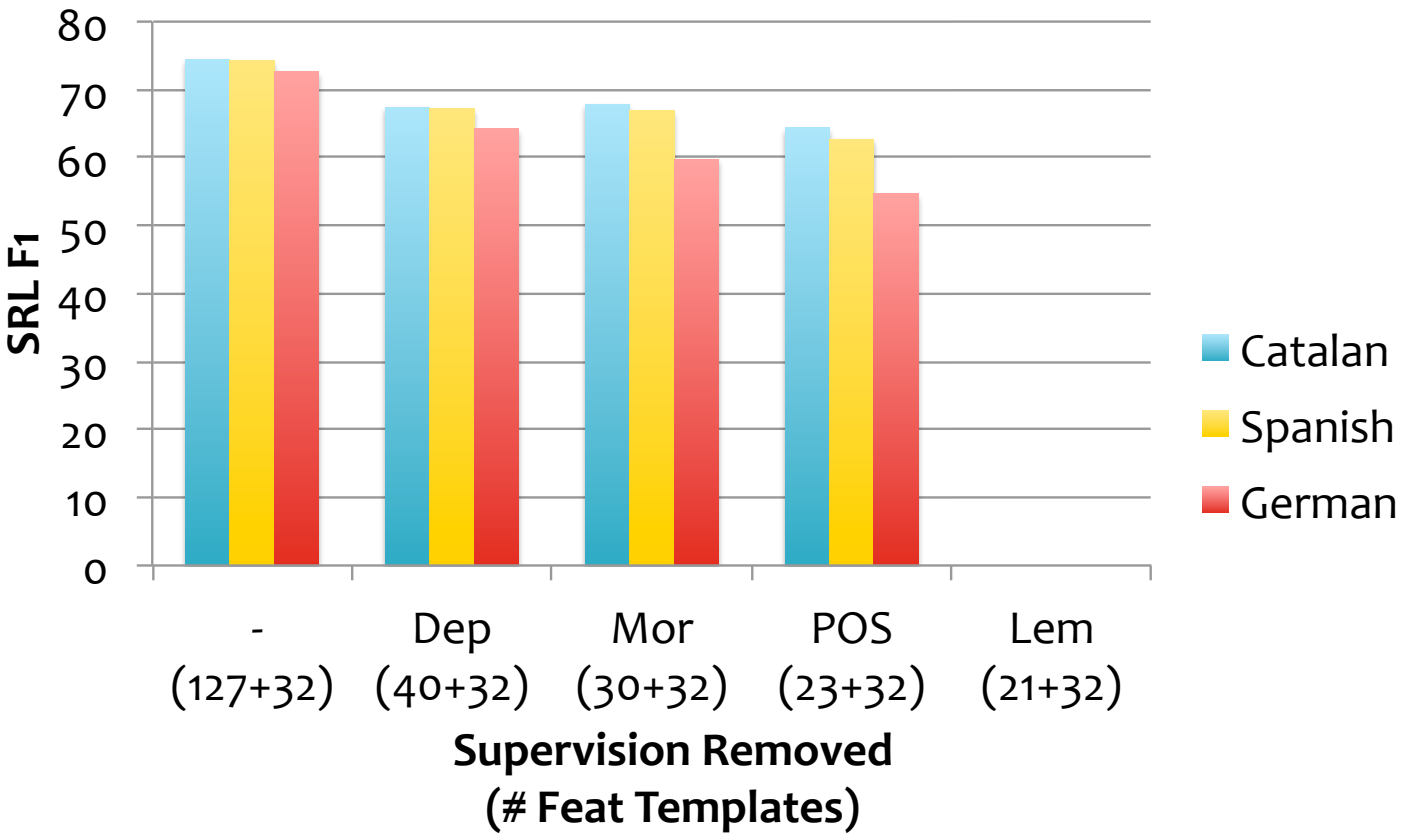
CoNLL-2009 Supervised Data
Semantic roles
~~Dependency parses~~
~~Morphology~~
Part-of-speech tags
Lemmas



Subtractive Experiments

Effectiveness of our joint models as the available supervision is decreased

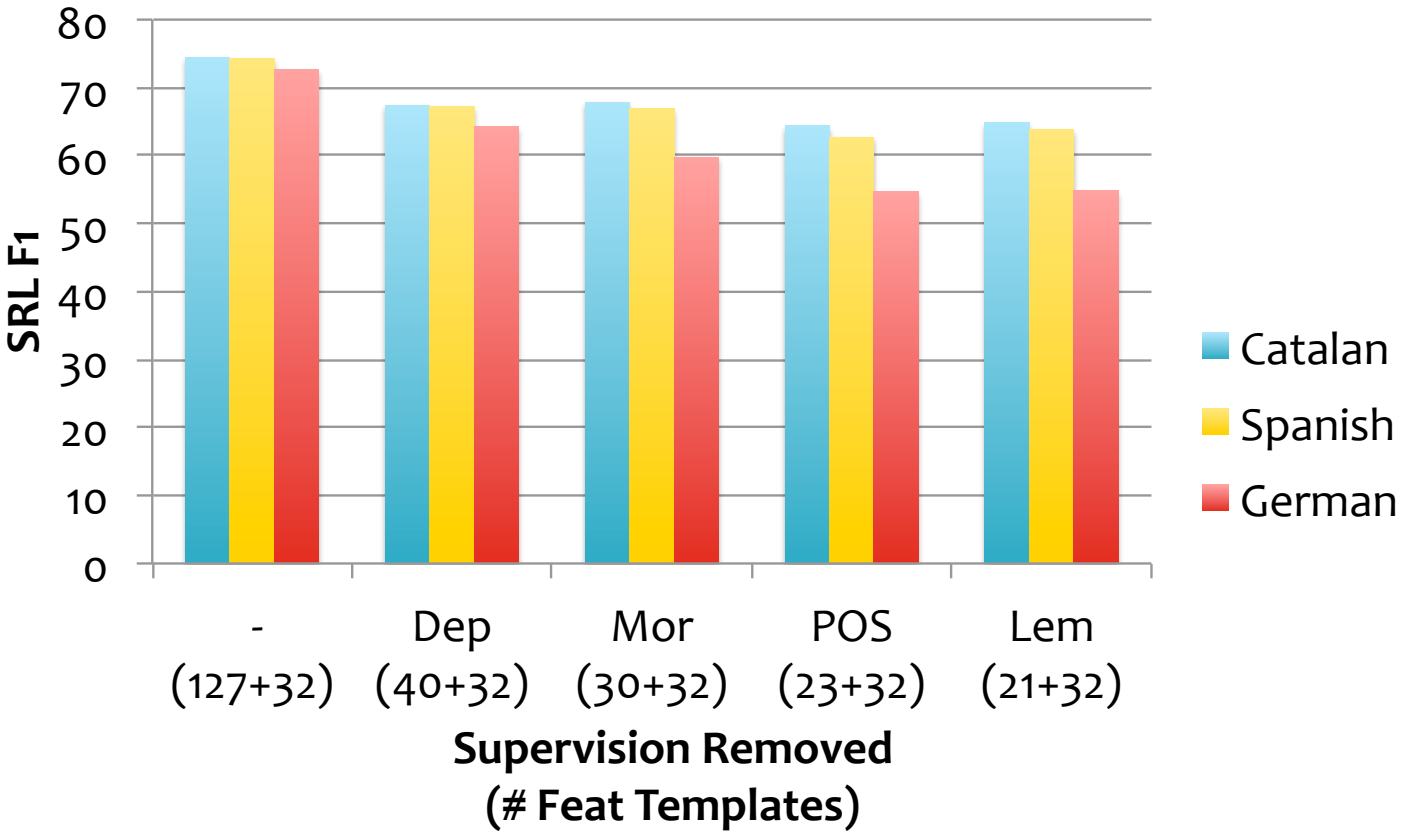
CoNLL-2009 Supervised Data
Semantic roles
~~Dependency parses~~
~~Morphology~~
~~Part-of-speech tags~~
Lemmas



Subtractive Experiments

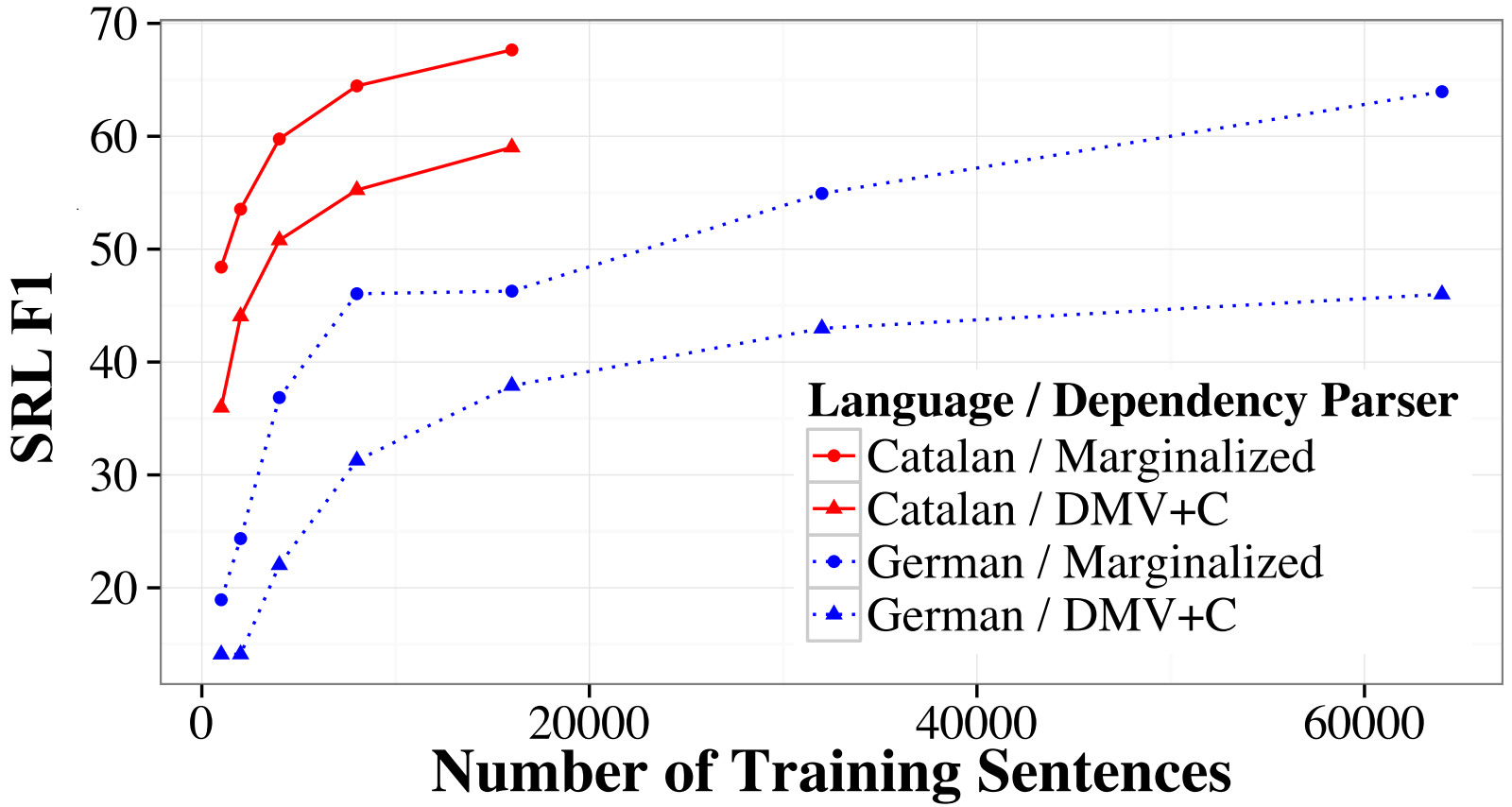
Effectiveness of our joint models as the available supervision is decreased

CoNLL-2009 Supervised Data
Semantic roles
~~Dependency parses~~
~~Morphology~~
~~Part-of-speech tags~~
~~Lemmas~~



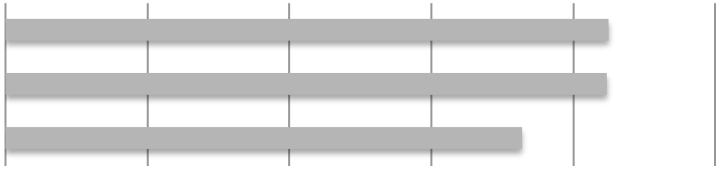
Learning Curves

Lowest resource setting: joint training yields higher SRL F1 than distant-supervision.



SRL Main Results

Parse	SRL Approach	Feat	Avg SRL F1
Gold	Pipeline	IG_C	**84.98**
	Pipeline	IG_B	84.74
	Naradowsky et al. (2012)		72.73

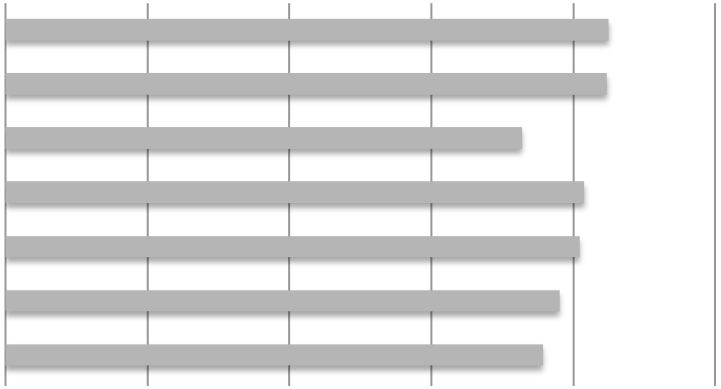


IG_C Features from Information Gain template selection, Coarse-Grained properties
IG_B Features from Information Gain template selection, Björkelund et al. (2009) properties

SRL Main Results

High-Resource

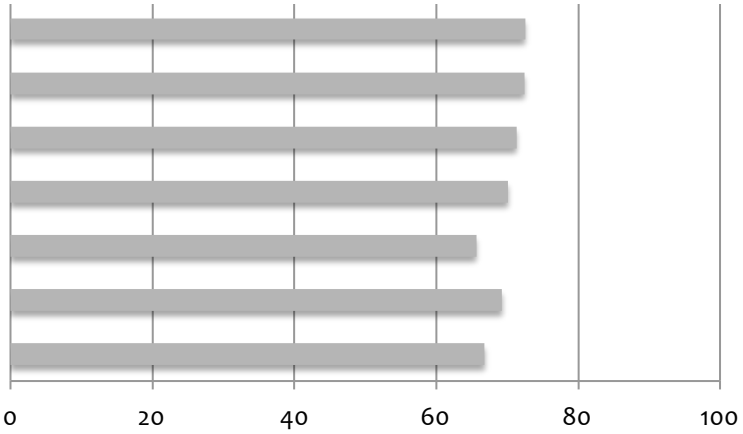
Parse	SRL Approach	Feat	Avg SRL F1
Gold	Pipeline	IG_C	**84.98**
	Pipeline	IG_B	84.74
	Naradowsky et al. (2012)		72.73
Super-vised	Björkelund et al. (2009)		**81.55**
	Zhao et al. (2009)		80.85
	Pipeline	IG_C	78.03
	Pipeline	IG_B	75.68



IG_C Features from Information Gain template selection, Coarse-Grained properties
IG_B Features from Information Gain template selection, Björkelund et al. (2009) properties

SRL Main Results

Parse	SRL Approach	Feat	Avg SRL F1
Low-Resource			
Joint (Margin-alized)	Joint	IG_C	**72.48**
	Joint	IG_B	72.40
	Naradowsky et al. (2012)		71.27
Distant (DMV+C)	Pipeline	IG_C	70.08
	Pipeline	IG_B	65.61
Unsup. (DMV)	Pipeline	IG_C	69.26
	Pipeline	IG_B	66.81

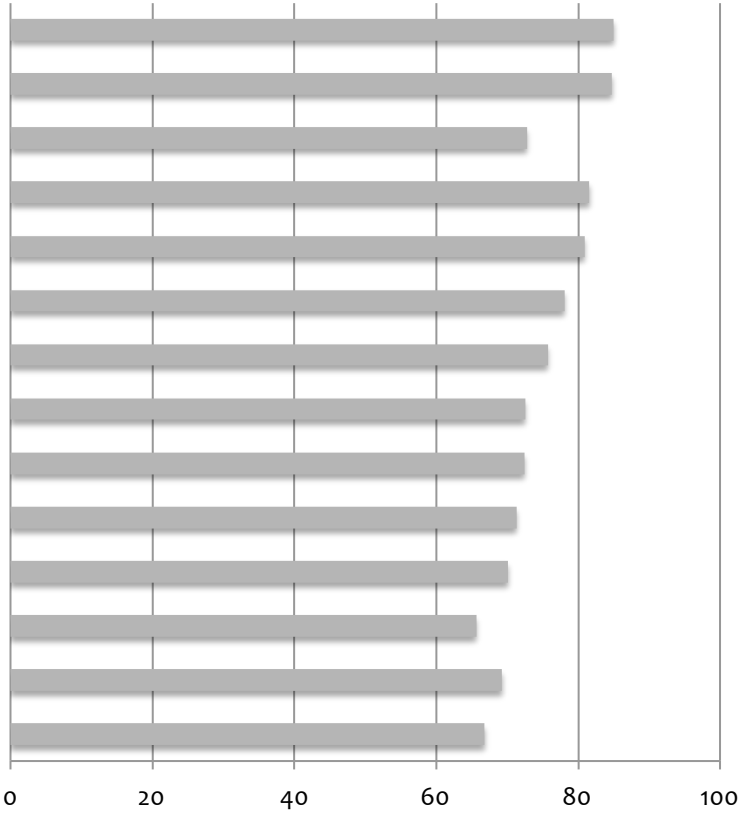


IG_C Features from Information Gain template selection, Coarse-Grained properties

IG_B Features from Information Gain template selection, Björkelund et al. (2009) properties

SRL Main Results

	Parse	SRL Approach	Feat	Avg SRL F1
High-Resource	Gold	Pipeline	IG_C	**84.98**
		Pipeline	IG_B	84.74
		Naradowsky et al. (2012)		72.73
	Super-vised	Björkelund et al. (2009)		**81.55**
		Zhao et al. (2009)		80.85
		Pipeline	IG_C	78.03
		Pipeline	IG_B	75.68
Low-Resource	Joint (Margin-alized)	Joint	IG_C	**72.48**
		Joint	IG_B	72.40
		Naradowsky et al. (2012)		71.27
	Distant (DMV+C)	Pipeline	IG_C	70.08
		Pipeline	IG_B	65.61
	Unsup. (DMV)	Pipeline	IG_C	69.26
		Pipeline	IG_B	66.81



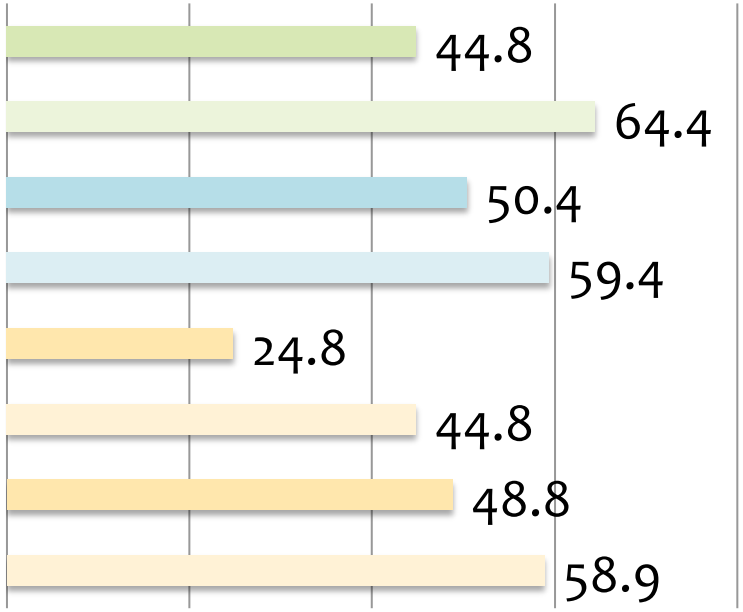
IG_C Features from Information Gain template selection, Coarse-Grained properties

IG_B Features from Information Gain template selection, Björkelund et al. (2009) properties

Comparison with Work in Grammar Induction in Low-Resource Setting

WSJ portion of Penn Treebank:

Approach	Distant Supervision	Unlabeled Syntactic Dependency Accuracy
Spitkovsky et al (2010)	None	44.8
Spitkovsky et al (2013)	None	64.4
Spitkovsky et al (2010)	HTML	50.4
Naseem and Barzilay (2011)	ACE05	59.4
DMV	None	24.8
DMV+C	SRL	44.8
Marginalized, IG_C	SRL	48.8
Marginalized, IG_B	SRL	58.9



- MBR decoding of marginalized grammars best DMV method
- May get gains with better search to break out of local optima

Is dependency accuracy the right evaluation metric?

In the joint model, higher **dependency accuracy (UAS)** does not *always* correlate with higher **Labeled F1 on SRL.**

Parse	SRL Approach	Feat	English UAS	English SRL F1
Joint (Margin-alized)	Joint	IG_C	44.2	76.16
	Joint	IG_B	52.6	75.57

Conclusions

- Semantic role labeling doesn't necessarily require a **long costly pipeline** of NLP tools

  (cf. Boxwell et al. (2011); Naradowsky et al. (2012))
- "Quality" of the **latent syntax** has a big effect

  (especially with limited end-task training data)
- Joint models seem to outperform the pipeline models in the low-resource setting

Questions?