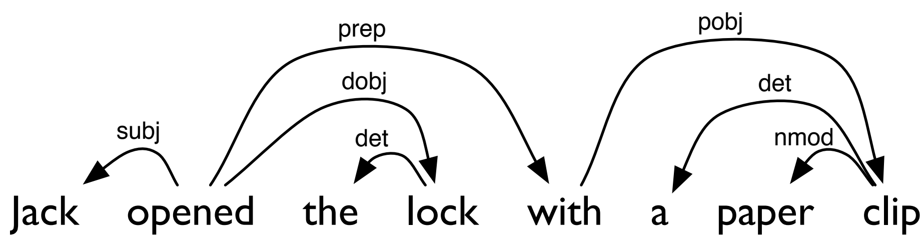
Semi-Supervised Semantic Role Labeling: Approaching from an Unsupervised Perspective

Ivan Titov and Alex Klementiev

From Syntax to Semantics

- Emergence of robust syntactic parsers [Collins 1999, Charniak 2001; Petrov and Klein 2006, McDonald 2005; Titov and Henderson 2007] for many languages has been one of the key successes of statistical NLP in recent years

- However, syntactic analyses are a long way from representing the meaning of sentences



Specifically, they do not define Who did What to Whom (and How, Where, When, Why, …)

- In other words, they do not specify the underlying predicate argument structure

Semantic Role Labeling (SRL)

Identification of arguments and their semantic roles

Jack opened the lock with a paper clip

PROTO-AGENT (A0)
an initiator/doer in the event [Who?]

PROTO-PATIENT (A1)
an affected entity [to Whom / to What?]

INSTRUMENT (A3)
the entity manipulated to accomplish the goal

Semantic Roles (PropBank-style)

Syntactic-Semantic Interface

Though syntactic and lexical representations are often predictive of the predicate argument structure, this relation is far from trivial, consider *alternations*:

John broke the window

The window broke

The window was broken by John

PROTO-AGENT (A0)
an initiator/doer in the event [Who?]

PROTO-PATIENT (A1)
an affected entity
[to Whom / to What?]

Approaches to SRL

- Supervised learning approaches (e.g., [Gildea and Jurafsky 2002; Johansson 2008])
  - Rely on large expert-annotated datasets (e.g., PropBank ~40k sentences)
  - Even then they provide very low coverage and are domain dependent
  - Annotated data is not available for many languages
- Semi-supervised methods – combine labeled and unlabeled data
  - Largely, extensions of supervised methods
  - Have relatively limited success so far – unannotated data adds little
- Unsupervised methods
  - E.g. [Lang and Lapata 2010, 2011; Titov and Klementiev 2012]

This work:

- Integrate labeled data into a state-of-the-art unsupervised system
- Compare performance of supervised/semi-supervised/unsupervised methods

Outline

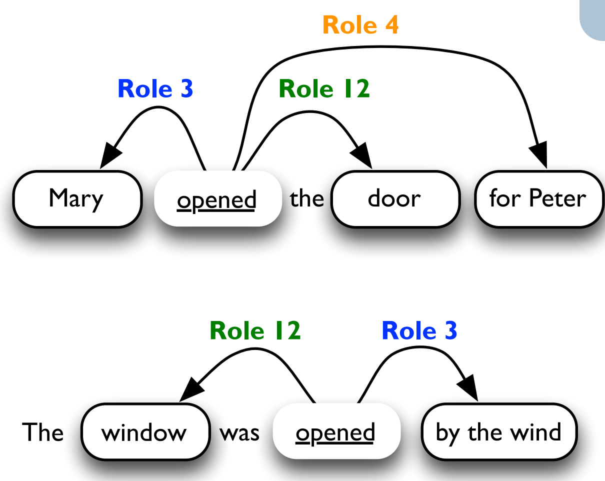
- Motivation
- Unsupervised semantic role induction
- Model and inference
  - Overview of the distance-dependent CRPs
  - A Bayesian model defining the process of joint generation of semantic, syntactic and lexical representations
- Semi-supervised extensions
  - Adding labeled data
  - Constructing informed priors
- Evaluation
  - Unsupervised, supervised and semi-supervised Results

Unsupervised Semantic Role Induction

- Semantic role labeling involves 2 sub-tasks:
  - Identification: identification of predicate arguments
  - Labeling: assignment of their sematic roles

Can be handled with heuristics (e.g. [Lang and Lapata, 2010])

Focus of this work



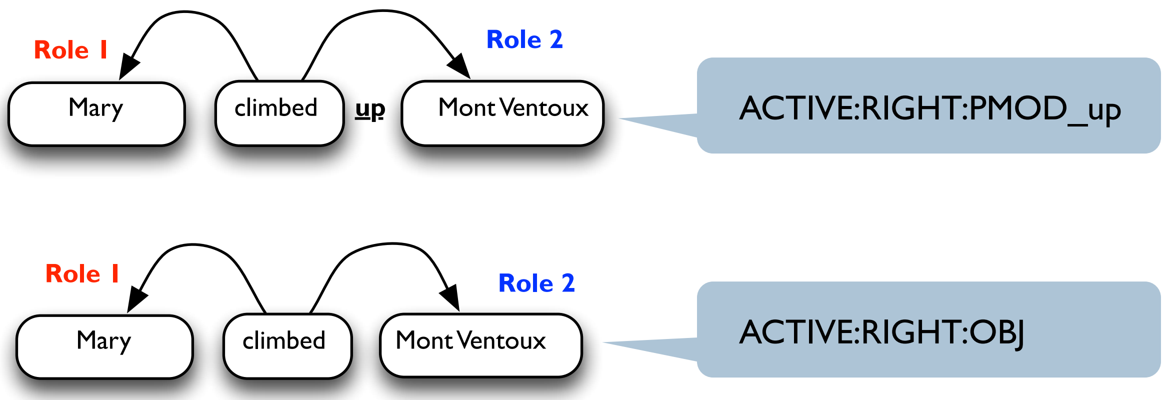
Goal: automatically induce semantic roles from unannotated data

- Assume that sentences are (auto-) annotated with syntactic trees
- Equivalent to clustering of argument occurrences (or "coloring" them)

Argument Keys

as in [Lang and Lapata, 2011]

- We identify arg occurrences with syntactic signatures (<u>argument keys</u>)
- E.g., some simple alternations like locative preposition drop



- Argument keys are designed to map mostly to a single role

We treat semantic role labeling as clustering of argument keys (instead of clustering arguments)

- E.g. in the example, we would cluster ACTIVE:RIGHT:OBJ and ACTIVE:RIGHT:PMOD_up

Our Goal

- Start with a state-of-the-art unsupervised model [Titov and Klementiev, 2012]

  Will call it *BayesSRL*

- The model induces clusterings of argument keys jointly across predicates
  - Intuition: clusterings are predicate specific, but similar, since many alternations are common across predicates (passivization, dativization, etc.)
  - The task is easier for some predicate than others
  - E.g., predicates *change* and *defrost* admit similar alternations but inducing it for *defrost* is easier: the set of possible argument fillers is more restricted

- Appropriate for semi-supervised setup
  - A reasonable approach should be able to propagate info across predicates

Our goal: extend the model to take advantage of labeled data

Outline

- Motivation
- Unsupervised semantic role induction
- Model and inference
  - Overview of the distance-dependent CRPs
  - A hierarchical Bayesian model defining the process of joint generation of semantic, syntactic and lexical representations
- Semi-supervised extension
  - Adding labeled data
  - Constructing informed priors
- Evaluation
  - Unsupervised, supervised and semi-supervised Results

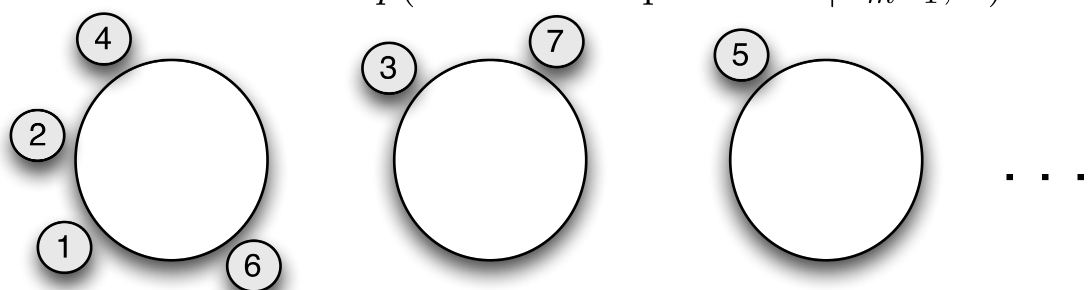
A Prior on the Partition of Argument Keys

- Can use CRP to define a prior on the partition of argument keys:
  - The first customer (argument key) sits the first table (role)
  - m-th customer sits at a table according to:

State of the restaurant once m-1 customers are seated

$$p(\text{previously occupied table } k | F_{m-1}, \alpha) \propto n_k$$
$$p(\text{next unoccupied table} | F_{m-1}, \alpha) \propto \alpha$$

Encodes rich-get-richer dynamics but not much more than that

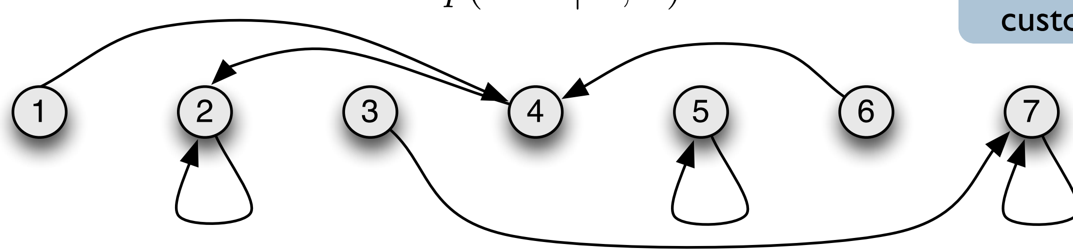


- An extension is distance-dependent CRP (dd-CRP):
  - m-th customer chooses a *customer* to sit with according to:

Entire similarity graph

$$p(\text{different customer } j | D, \alpha) \propto d_{m,j}$$
$$p(\text{itself} | D, \alpha) \propto \alpha$$

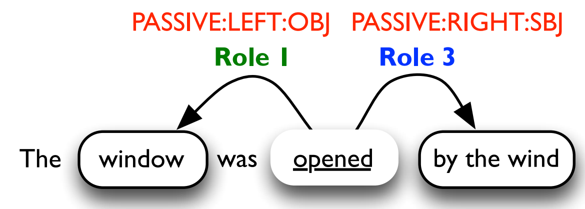
Similarity between customers m and j



A Prior on the Partition of Argument Keys

- Similarity graph *D* couples distinct but similar clusterings of argument keys across predicates
  - Vertices are argument keys
  - Weights are similarity scores for each pair of argument keys
- We treat *D* as a latent random variable drawn from a prior over weighted graphs
  - First drawn from a prior
  - Used to generate each of the clusterings for every predicate
- We induce *D* automatically within the model
  - This is in contrast to all the previous work on dd-CRP where similarities were used to encode prior knowledge

Generative Story



At least one argument

Draw first argument

Continue generation

Draw more arguments

$$
\begin{aligned}
&for\ each\ predicate\ p = 1, 2, \cdots :\\
&\quad for\ each\ occurrence\ l\ of\ p :\\
&\quad\quad for\ every\ role\ r \in B_p :\\
&\quad\quad\quad if\ [n \sim Unif(0,1)] = 1 :\\
&\quad\quad\quad\quad \mathbf{GenArgument}(p, r)\\
&\quad\quad\quad while\ [n \sim \psi_{p,r}] = 1 :\\
&\quad\quad\quad\quad\quad \mathbf{GenArgument}(p, r)
\end{aligned}
$$

$\mathbf{GenArgument}(p, r)$

$$
\begin{aligned}
&k_{p,r} \sim Unif(1, \ldots, |r|)\\
&x_{p,r} \sim \theta_{p,r}
\end{aligned}
$$

Draw argument key

Draw argument filler

$$
\begin{aligned}
&D \sim NonInform\\
&\text{for each predicate } p = 1, 2, \ldots :\\
&\quad\quad B_p \sim dd\text{-}CRP(\alpha, D)
\end{aligned}
$$

$$
\begin{aligned}
&\text{for each predicate } p = 1, 2, \ldots :\\
&\quad\quad \text{for each role } r \in B_p\text{:}\\
&\quad\quad\quad\quad \theta_{p,r} \sim DP(\beta, H^{(A)})\\
&\quad\quad\quad\quad \psi_{p,r} \sim Beta(\eta_0, \eta_1)
\end{aligned}
$$

Inference: iterate between clustering given similarity graph *D*, and re-estimating *D* (see [Titov and Klementiev 2012])

Outline

- Motivation
- Unsupervised semantic role induction
- Model and inference
  - Overview of the distance-dependent CRPs
  - A Bayesian model defining the process of joint generation of semantic, syntactic and lexical representations
- Semi-supervised extension
  - Adding labeled data
  - Constructing informed priors
- Evaluation
  - Unsupervised, supervised and semi-supervised Results

Exploiting Labeled Data

- Idea I: integrate labeled data into a generative model
  - Maximize joint likelihood of the observed data (i.e. clamp the observed labels)
  - BayesSRL makes hard clustering decisions

Problem: imperfect purity of arg keys + potential annotation errors may mean no possible clusterings may be compatible with labeled data

  - Change generative story: with small probability ϵ, draw a random argument key

GenArgument(p, r)

$$
\begin{aligned}
& b \sim Bernoulli(\epsilon) \\
& if\ b = 1: \\
& \quad k_{p,r} \sim H^{(K)} \\
& else: \\
& \quad k_{p,r} \sim Unif(1, \ldots, |r|) \\
& \quad x_{p,r} \sim \theta_{p,r}
\end{aligned}
$$

Noisy arg key

Draw arg key

Draw arg filler

Exploiting Labeled Data

- Idea II: use labeled data to construct an informed prior over argument key clusterings
- We estimate:
  - How likely two arg keys k and k' are in the same role
  - How likely a specific key k is to be left unclustered

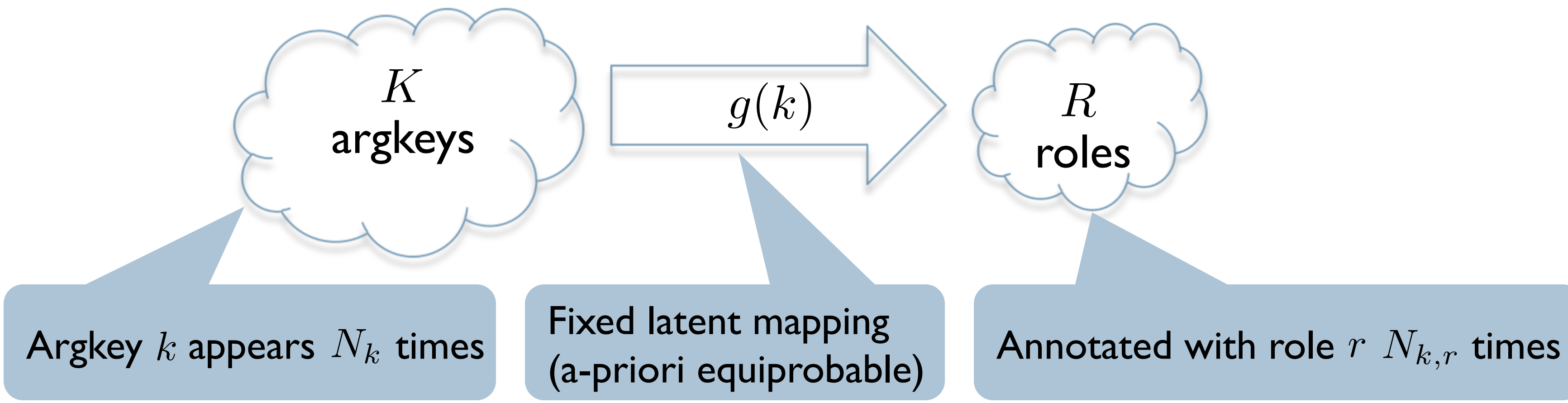
Use to set similarity $\hat{d}_{k,k'}$

… and concentration parameter for dd-CRP

Note: this is not model estimation but an extrinsic way to set priors

Exploiting Labeled Data to Construct Informed Priors

- Consider a single predicate:



- When generating a label, choose any other $R-1$ roles with small prob. γ
- Thus, the probability of labeled examples (role assignments) X_k for key k is:

$$P(X_k | g(k) = r) = (1-\gamma)^{N_{k,r}} \left(\frac{\gamma}{R-1} \right)^{N_k - N_{k,r}}$$

Exploiting Labeled Data to Construct Informed Priors

- The probability of labeled examples (role assignments) X_k for key k is:

$$P(X_k|g(k) = r) = (1-\gamma)^{N_{k,r}} \left(\frac{\gamma}{R-1}\right)^{N_k - N_{k,r}}$$

- The joint probability of two sets of labels X_k and $X_{k'}$

$$P(X_k, X_{k'}|g(k) = g(k'))$$
$$= \sum_r P(X_k|g(k) = r)P(X_{k'}|g(k') = r)$$

Same role (any)

$$P(X_k, X_{k'}|g(k) \neq g(k'))$$
$$= \sum_r P(X_k|g(k) = r) \sum_{r' \neq r} P(X_{k'}|g(k') = r')$$

Different role

The posterior that two keys belong to the same role $P(g(k) = g(k')|X)$ is given by re-normalizing the above expressions

Exploiting Labeled Data to Construct Informed Priors

- In dd-CRP, $\hat{d}^{(p)}_{kk'}$ encodes how much more likely k and k' are clustered together than by random chance, so we compute it as:

$$\hat{d}^{(p)}_{kk'} = \frac{P(g(k) = g(k')|X)}{P(g(k) = g(k'))}$$

Posterior prob. that two argkeys are in the same role

Prior that two argkeys are in the same role = $1/R$

- Insufficient for infrequent (most) predicates, so we also compute $\hat{d}_{kk'}$ *across* predicates
- When generating partitions B_p, we multiply $\hat{d}^{(p)}$, $\hat{d}$ and automatically induced d
- The other dd-CRP parameter $\hat{\alpha}^{(p)}_k$ can be computed similarly

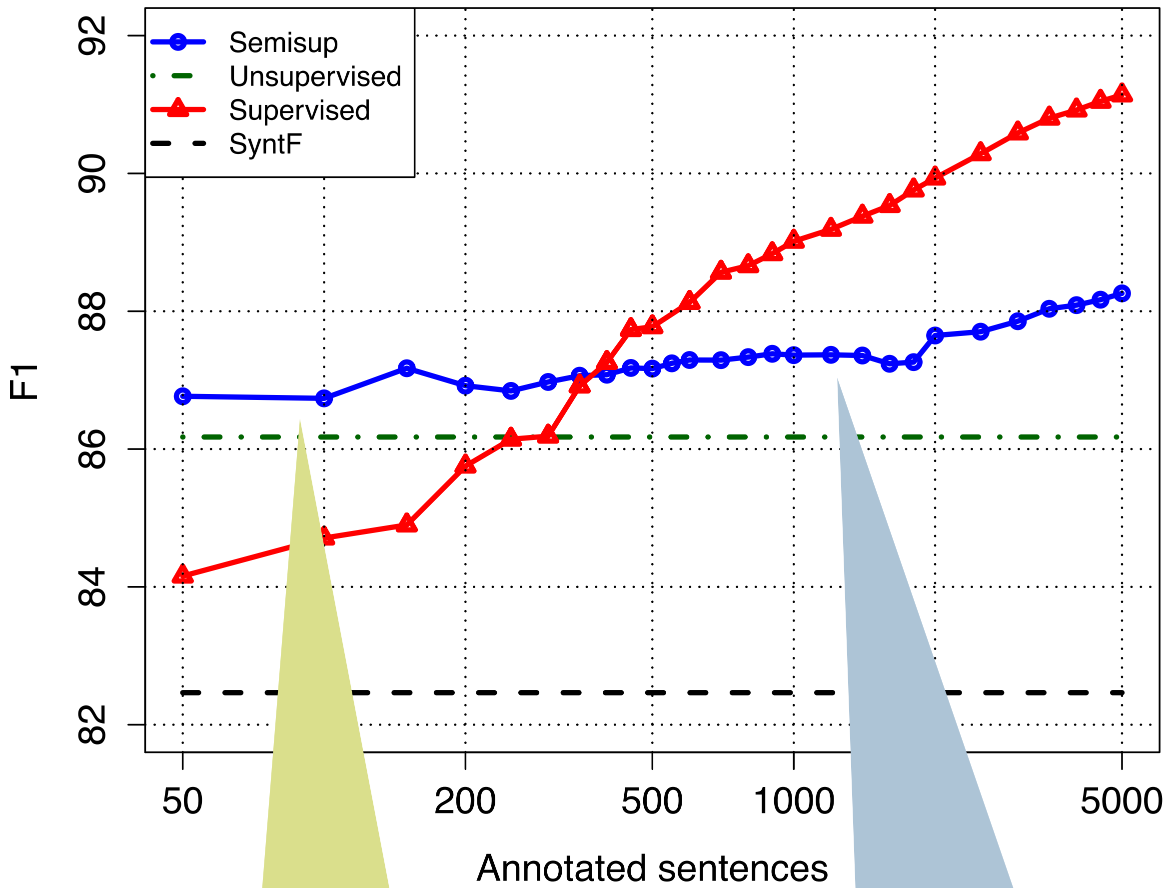
Outline

- Motivation
- Unsupervised semantic role induction
- Model and Inference
  - Overview of the distance-dependent CRPs
  - A Bayesian model defining the process of joint generation of semantic, syntactic and lexical representations
- Semi-supervised extension
  - Adding labeled data
  - Constructing informed priors
- Evaluation
  - Unsupervised, supervised and semi-supervised Results

Benchmark Dataset: PropBank (CoNLL 09)

- Train on one half (20,000 sentences) of the dataset, evaluate on the other
- Annotate with predicted dependencies [Johansson and Nugues, 2008]
- Select non-auxiliary verbs as predicates, identify arguments using the heuristic of [Lang and Lapata, 2011]
- Evaluate argument labeling stage using standard clustering measures: Purity, Collocation (and F1) and Homogeneity, Completeness (and V-Measure)
- Compare with *Unsupervised* [Titov and Klementiev, 2012], *Supervised* [Johansson and Nugues, 2008] and *SyntF* (syntactic function)

Argument Labeling Evaluation

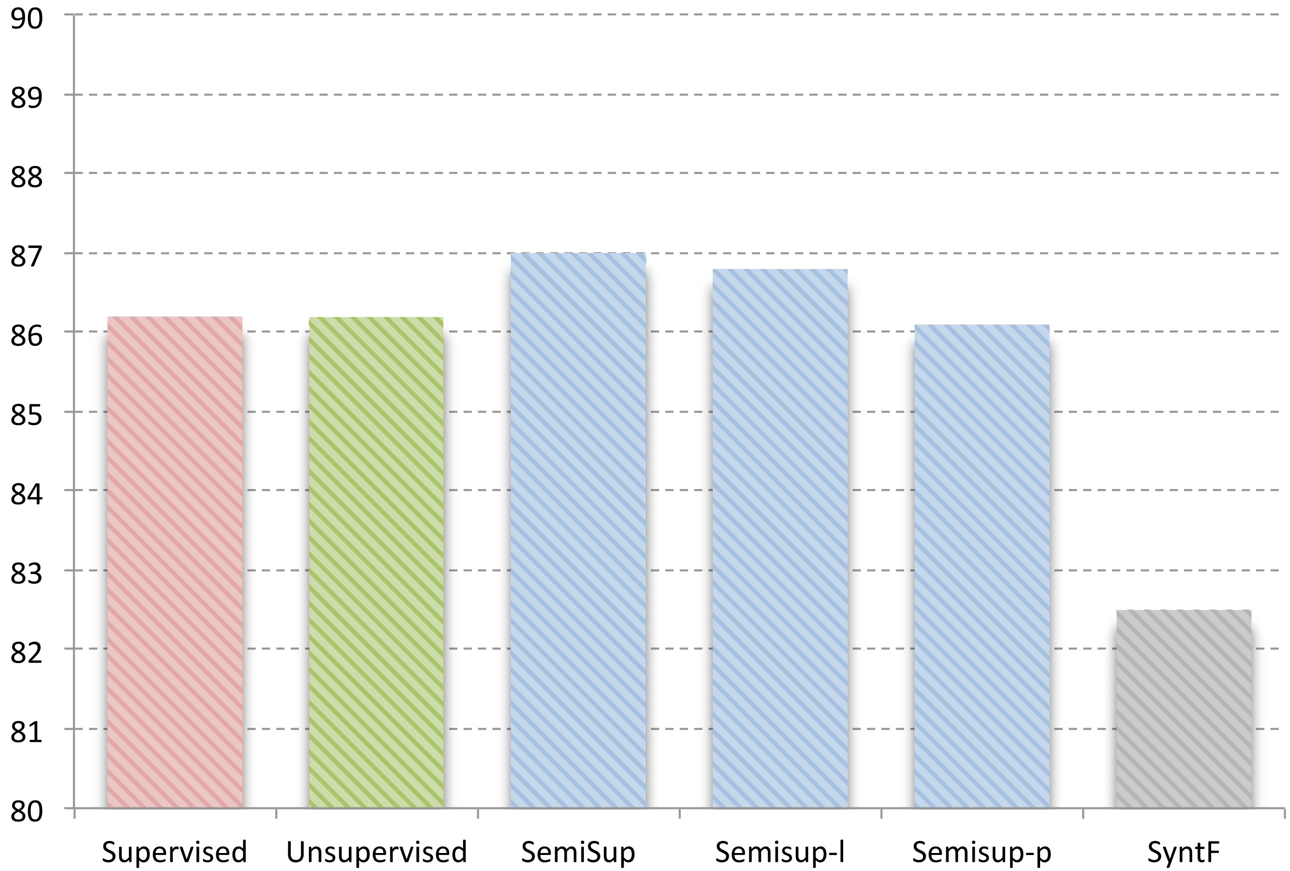


Effective for low resource setting

Coarse granularity means lower asymptotic performance

Argument Labeling Evaluation

One point on the curve: 300 labeled sentences



Semusup

without labeled data in gen story

without informed prior

Conclusions

- Demonstrated that unsupervised techniques can be improved by eploiting a small amount of annotated data

Results competitive with supervised approaches in low resource setting

- Uncovered deficiencies of unsupervised approaches

Overly coarse modeling of syntax-semantics interface results in lower asymptotic performance