A Bayesian Approach to Unsupervised Semantic Role Induction

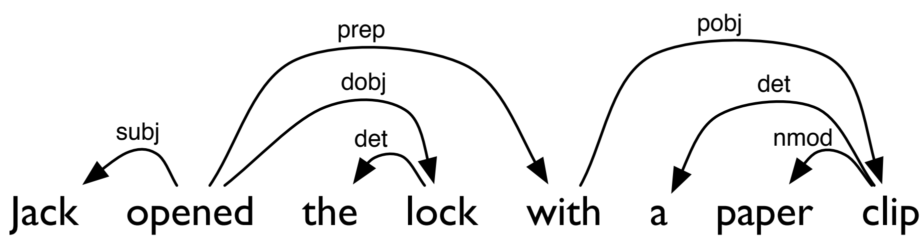
Ivan Titov and Alex Klementiev

CLUSTER OF EXCELLENCE

UNIVERSITÄT DES SAARLANDES

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 <u>their semantic roles</u>
- Example: predicate *open*

Jack opened the lock with a paper clip

Semantic Roles (PropBank-style):

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

Syntactic-Semantic Interface

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

(1) John broke the window

(2) The window broke

(3) The window was broken by John

Semantic Roles:

AGENT – an initiator/doer in the event [Who?]

PATIENT - 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
  - Have relatively limited success so far (e.g., Furstenau and Lapata [09]; Deschacht and Moens [09])
- Unsupervised methods
  - This work, also Lang and Lapata [2010, 2011] and Grenager and Manning [2006]

Our main contributions:

- a Bayesian model of unsupervised SRL, substantially outperforming previous work
- Induction of a representation encoding alternation patterns shared across predicates

Outline

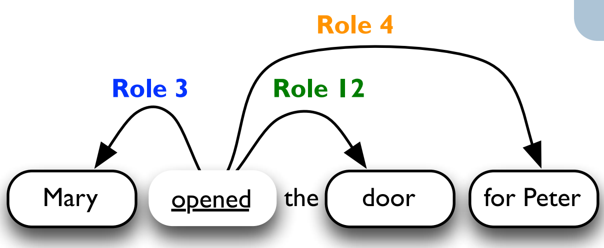
- Task and Approach Overview
  - Semantic role induction without labeled data
- Model and Inference
  - Overview of the distance-dependent CRPs
  - A hierarchical Bayesian model defining the process of joint generation of semantic, syntactic and lexical representations
- Evaluation
  - Results on a human-annotated corpus

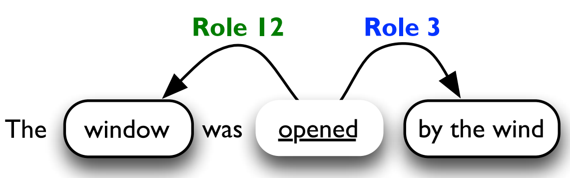
Our task

- 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



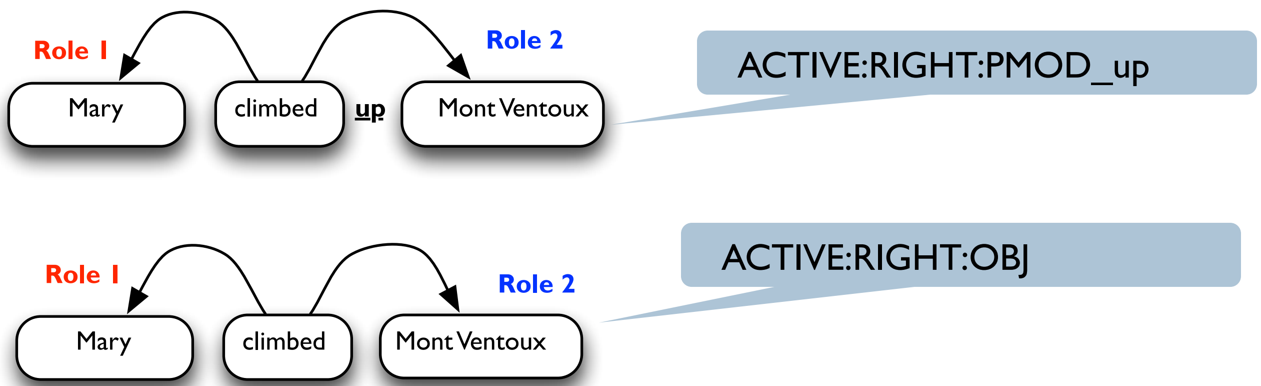


Our goal: induce semantic roles automatically from unannotated texts

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

Argument Keys

- We identify arg occurrences with syntactic signatures (<u>argument keys</u>) (as in Lang and Lapata [2011])
- E.g., some simple alternations like locative preposition drop



- Argument keys are designed so that to map mostly to a single role
- Instead of clustering occurrences we cluster argument keys
- Here, we would cluster ACTIVE:RIGHT:OBJ and ACTIVE:RIGHT:PMOD_up together
  - More complex alternations require multiples pairs of arg keys clustered

Factored Model

- Our first model (**Factored)** clusters argument keys for every predicate in isolation.

- These clusterings
  - are different as verbs admit different alternations
  - but expected to be similar: many alternations are common and licensed by many predicates (passivization, dativization, etc)

Coupled Model

- Consequently, propose an extension (**Coupled**) to induce the clusterings jointly
  - Do not split the learning data
  - 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

- This is done by inducing a similarity score for every pairs of argument keys
  - Similarities are <u>learned</u>, rather than specified by hand, as part of the inference process

Signals for Semantic Role Induction

- Selection preferences:
  - Two argument signatures are likely to correspond to the same role if the corresponding sets of arguments are similar.

- Duplicate roles are unlikely to occur. E.g. this coloring is a bad idea:

  John taught students math

- Predicates admit similar alternation patterns (reuse them)

How to encode this in a statistical model?

Outline

- Task and Approach Overview
  - Semantic role induction without labeled data
- 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
- Evaluation
  - Results on a human-annotated corpus

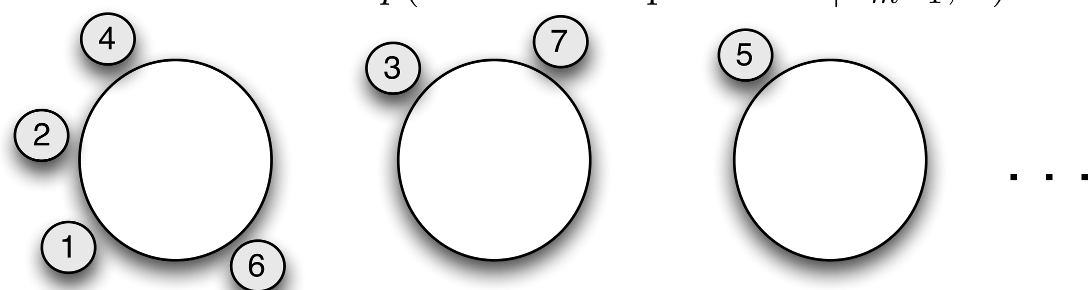
A Prior on the Partition of Argument Keys

- 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

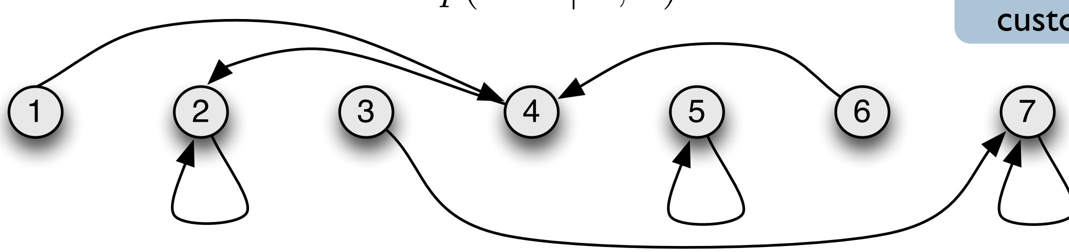


- An extension is distance-dependent CRP (dd-CRP):
  - m-th customer chooses a *customer* to sit with according to:

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

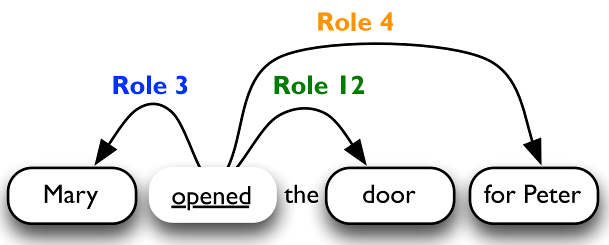


A Prior on the Partition of Argument Keys

- Similarity graph *D* to 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

Bayesian Induction of Semantic Roles

- Given a (large) collection of sentences annotated with (transformed) syntactic dependencies $\{x_i\}_{i=1}^n$
- We want to induce semantic roles $\{m_i\}_{i=1}^n$



- Define a family of generative models $P(m, x|\boldsymbol{\theta})$ encoding our assumptions
- In the prior probability $P(\boldsymbol{\theta})$ over parameters $\boldsymbol{\theta}$, we encode our beliefs
- We incorporate latent variables $\boldsymbol{z}$ (our latent weighted graph *D*)
- We want to find the maximum-a-posteriori clustering given the observable data

$$\{\hat{m}_i\}_{i=1}^n = \arg\max \int \prod_{i=1}^n P(m_i, x_i, z_i|\boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{z}$$

Model parameters

(1) For roles, the distribution over argument fillers is sparse

- We use a sparse prior, Hierarchical Dirichlet Processes [Teh et al, 05]

(2) Each predicate undergoes a small number of alternations

- We use sparse Dirichlet priors to encode the linking

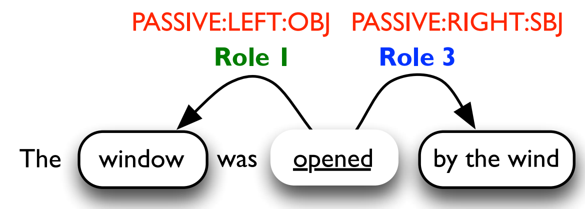
(3) The same semantic role rarely appears twice

- Use a non-symmetric Dirchlet prior for the corresponding geom. distrib

(4) Argument key clusterings for different predicates are related

- Induce a shared weighted graph used in a (distance-dependent) Chinese Restaurant Process [Blei and Frazer 11] prior for each clustering

Generative Stories for Factored and Coupled Models



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

At least one argument

Draw first argument

Continue generation

Draw more arguments

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

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

Draw argument key

Draw argument filler

Factored model:

for each predicate $p = 1, 2, \ldots$:

$\quad B_p \sim CRP(\alpha)$

Coupled model:

$D \sim NonInform$

for each predicate $p = 1, 2, \ldots$:

$\quad B_p \sim dd\text{-}CRP(\alpha, D)$

for each predicate $p = 1, 2, \ldots$:

$\quad$ for each role $r \in B_p$:

$$
\begin{aligned}
\theta_{p,r} &\sim DP(\beta, H^{(A)}) \\
\psi_{p,r} &\sim Beta(\eta_0, \eta_1)
\end{aligned}
$$

Inference

$$\{\hat{m}_i\}_{i=1}^n = \arg\max_{\{m_i\}_{i=1}^n} \int \prod_{i=1}^n P(m_i, x_i|\boldsymbol{\theta}) P(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

- We use approximate maximum a-posteriori (MAP) decoding to induce semantic representations
  - Similar techniques has been used in the context of Dirichlet process mixtures
- An EM-like inference algorithm for the Coupled model:
  - Start with uniform similarities
  - Iterate between
    - Inducing new clusterings ***m*** of argument keys for each predicates given the similarity graph ***D***
    - Reestimate the similarity graph ***D***

Outline

- Task and Approach Overview
  - Semantic role induction without labeled data
- 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
- Evaluation
  - Results on a human-annotated corpus

Benchmark Dataset: PropBank (CoNLL 08)

- Evaluation of semantic role induction

- Purity measures the degree to which each induced role contains arguments sharing the same gold ("true") role

$$PU = \frac{1}{N} \sum_i \max_j |G_j \cap C_i|$$

Gold role

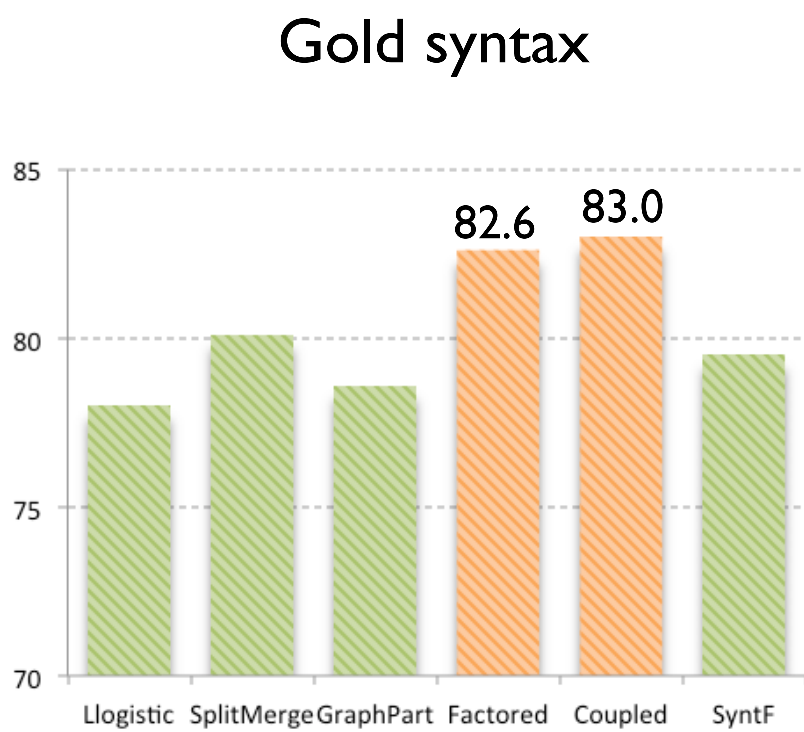
Induced role

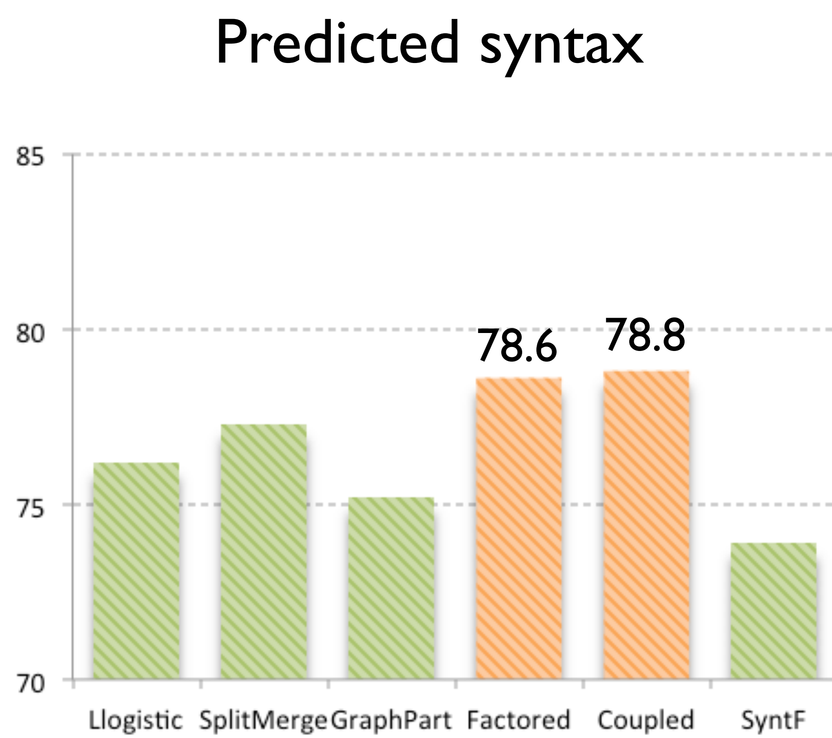
- Collocation evaluates the degree to which arguments with the same gold roles are assigned to a single induced role

$$CO = \frac{1}{N} \sum_j \max_i |G_j \cap C_i|$$

- Report F1, harmonic mean of PU and CO

PropBank (CoNLL 08) with Gold Argument ID





State-of-the-art methods

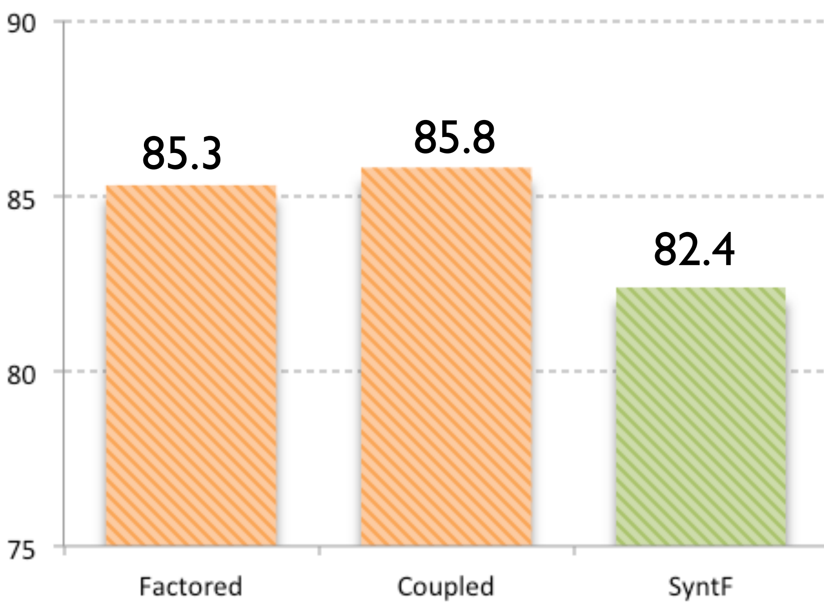
Our models

Our models

Deterministic mapping from syntactic relations

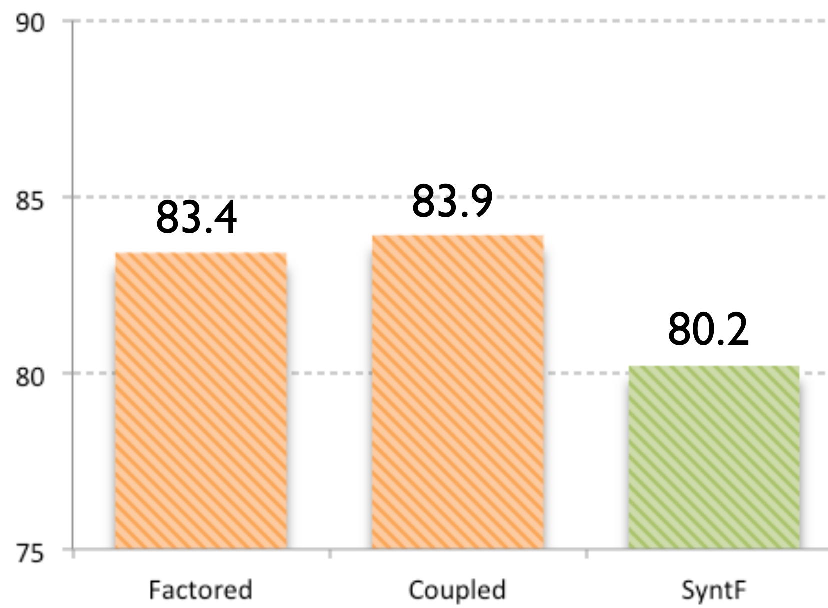
PropBank (CoNLL 08) with Predicted Argument ID

Gold syntax



Our models

Predicted syntax



Our models

Benchmark Dataset: PropBank (CoNLL 08)

Looking into induced graph encoding 'priors' over clustering arguments keys, the most highly ranked pairs encode (or partially encode)

Encoded as (ACTIVE:RIGHT:OBJ_if, ACTIVE:RIGHT:OBJ_whether)

- Passivization
- Near-equivalence of subordinating conjunctions and prepositions
  - E.g., *whether* and *if*
- Benefactive alternation

  Martha carved a doll for the baby

  Martha carved the baby a doll
- Dativization

  I gave the book to Mary

  I gave Mary the book
- Recovery of unnecessary splits introduced by argument keys

Conclusions

- We proposed a Bayesian model for unsupervised SRL
- Best reported scores on PropBank
- First to induce alternation patterns shared across predicates

- The proposed multi-task clustering approach is a general method
  - Can be used as a component in many Bayesian models for NLP and beyond

- The data, code and evaluation scripts will be available on our web-pages within a week or two.