Unsupervised and Cross-lingual Induction of Semantic Representations

Ivan Titov

Joint work with Alex Klementiev,
Mike Kozhevnikov, Ashutosh Modi and Binod Bhattarai

Universiteit van Amsterdam

Why semantic representations?

Question Answering about knowledge in a collection of biomedical publications:

Question: What does cyclosporin A suppress?

Answer: expression of EGR-2

Sentence: As with EGR-3 , expression of EGR-2 was blocked by cyclosporin A .

Question: What inhibits tnf-alpha?

Answer: IL -10

Sentence: Our previous studies in human monocytes have demonstrated that interleukin (IL) -10 inhibits lipopolysaccharide (LPS) -stimulated production of inflammatory cytokines , IL-1 beta , IL-6 , IL-8 , and tumor necrosis factor alpha by blocking gene transcription .

We need to abstract away from specific syntactic and lexical realizations

Why cross-lingual semantic representations?

- Improvements for individual languages
  - Crosslingual learning has been successful in syntax [Kuhn, 2004; Snyder et. al., 2009; McDonald et al., 2011] and morphology [Snyder and Barzilay, 2008]
  - Should be even more beneficial for inducing semantics, as semantics is generally better preserved in translation

Crosslingual (unknown) regularities provide a signal for learning

Can encode directly to drive learning: e.g. one-to-one correspondences between semantic representations

- Induced semantic relationships across multiple languages
  - Immediately useful for multilingual problems such as machine translation, multilingual web search, annotation projection across languages, …

Outline

- Induction of events and their participants
  - unsupervised models of semantic roles
  - joint induction of frames and roles
  - cross-lingual extension and comparison with projection and transfer
- Induction of semantic representations of words (and phrases)
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, lexicon induction)

Representing events and their participants

- A semantic frame [Fillmore 1968] is a conceptual structure describing a situation, object, or event along with associated properties and participants

- Example: CLOSURE / OPENING frame

  Jack opened the lock with a paper clip

  Semantic Roles (aka Frame Elements):

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

  PATIENT - an affected entity [to Whom / to What?]

  INSTRUMENT – the entity manipulated to accomplish the goal

  Other roles for CLOSURE/OPENING frame: BENEFICIARY, FASTENER, DEGREE, CIRCUMSTANCES, MANIPULATOR, PORTAL, …

Syntax-Semantics Interface

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

(1) John broke the window

(2) The window broke

(3) The window was broken by John

(4) John busted the window

(5) The window was destroyed by John

(6) John tore down the window

Alternations

The same relation is encoded by different predicates (incl. a multiword expression)

Semantic Roles:

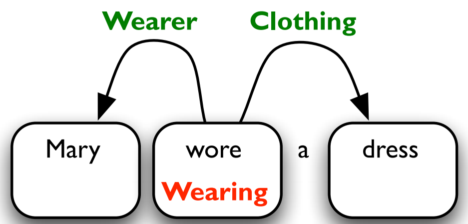
AGENT – an initiator/doer in the event [Who?]

PATIENT - an affected entity [to Whom / to What?]

Supervised learning of semantic representations is challenging: datasets provide low coverage, are domain-specific and available only for a few languages

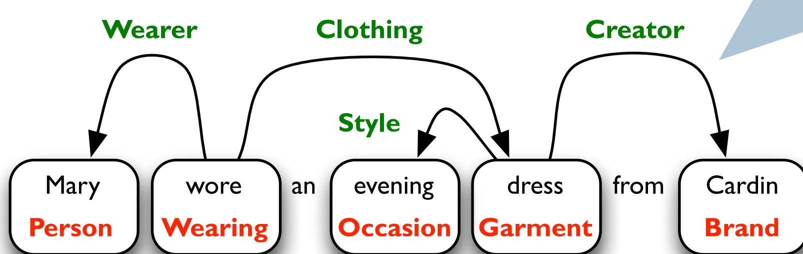
Our task

- Semantics is encoded by semantic dependency graphs [Johansson, 2008]

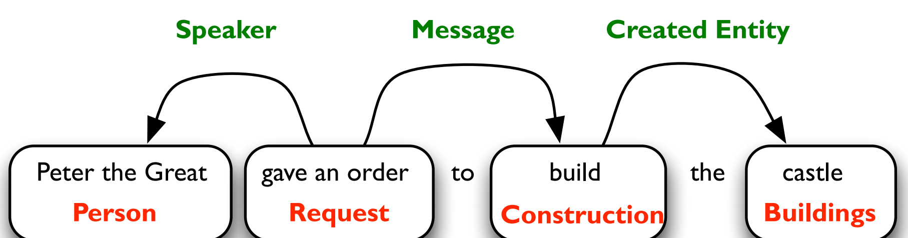


- Arguments often evoke their own frames

For simplicity we assume that all of them evoke frames



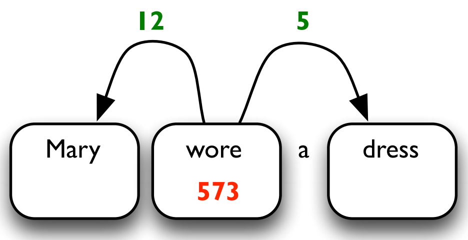
- Arguments and predicates often expressed by multiword expressions



Induce these representations automatically from unannotated texts

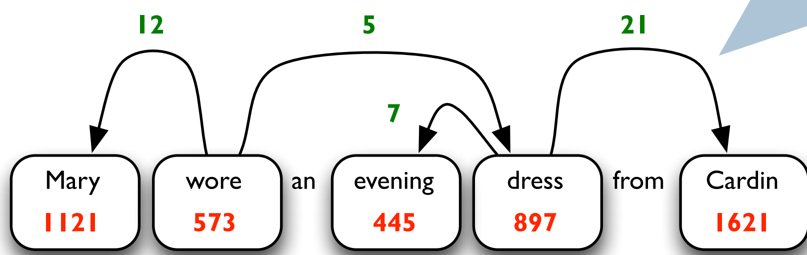
Our task

- Semantics is encoded by semantic dependency graphs [Johansson, 2008]

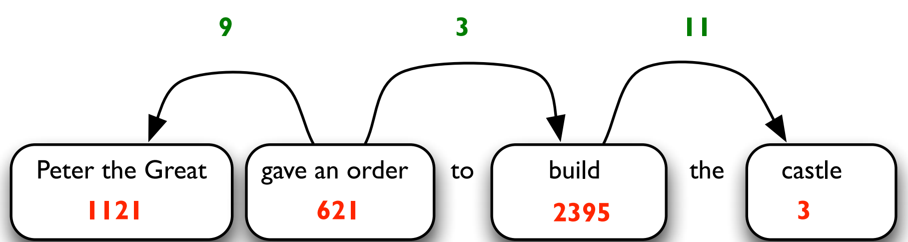


For simplicity we assume that all of them evoke frames

- Arguments often evoke their own frames



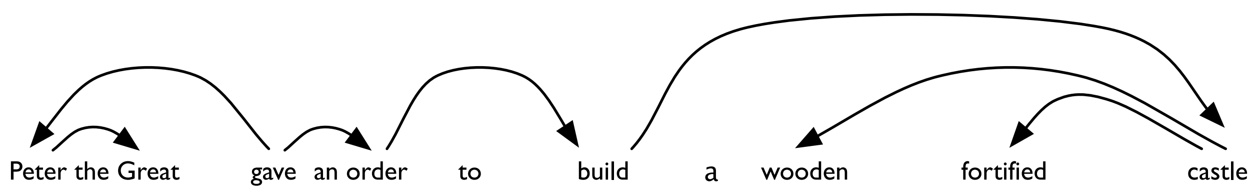
- Arguments and predicates often expressed by multiword expressions



Induce these representations automatically from unannotated texts

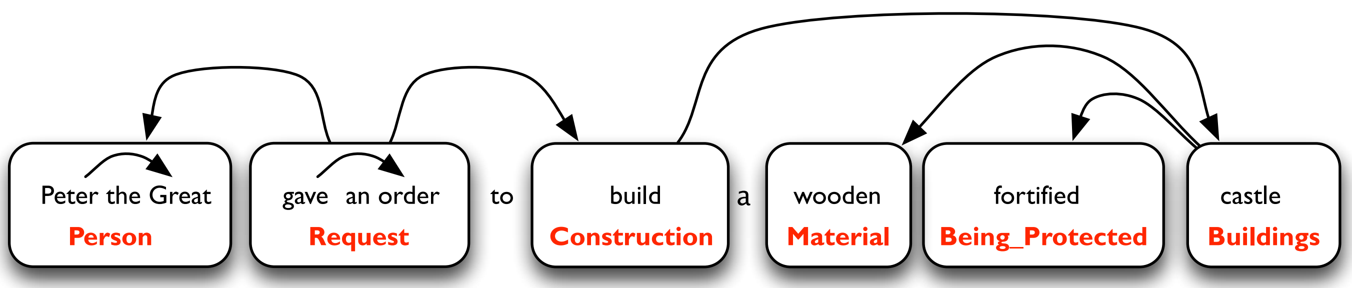
Induction of Frame-Semantic Information

- The semantic induction task involves 3 sub-tasks
  - Construction of a transformed syntactic dependency graph (~ argument identification)



Induction of Frame-Semantic Information

- The semantic induction task involves 3 sub-tasks
  - Construction of a transformed syntactic dependency graph (~ argument identification)
  - Induction of frames (and clusters of arguments)

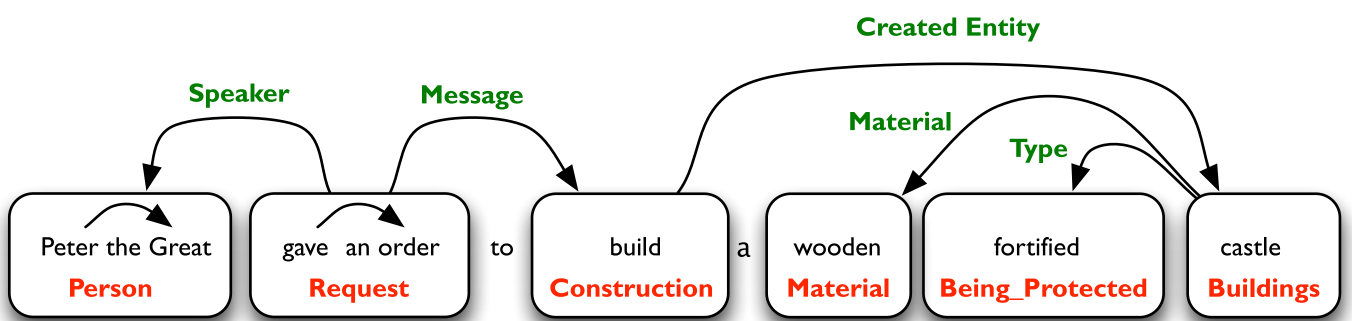


Induction of Frame-Semantic Infor

Handled with a simple heuristic or a simple classifier

- The semantic induction task involves 3 sub-tasks
  - Construction of a transformed syntactic dependency graph (~ argument identification)
  - Induction of frames (and clusters of arguments)
  - Role Induction

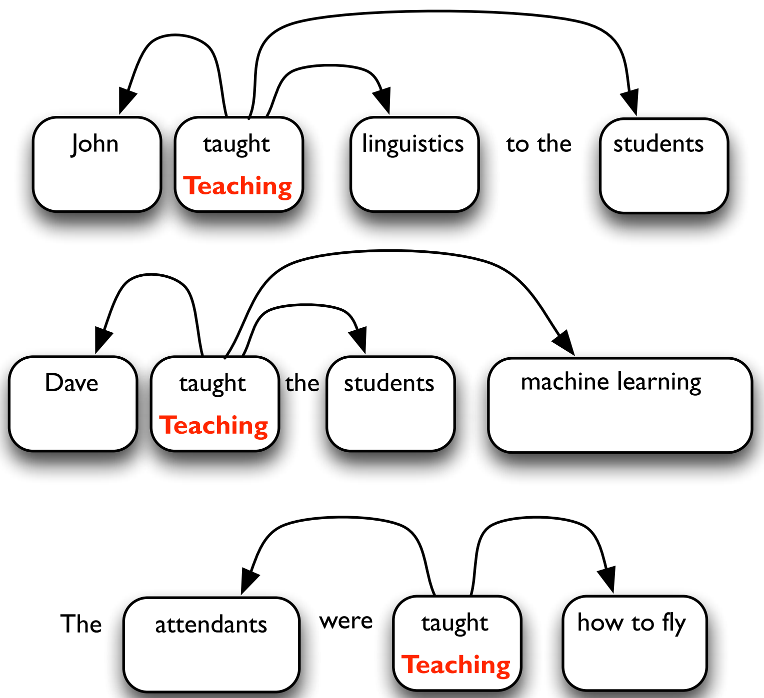
We model these sub-tasks jointly within our Bayesian model



Different from much of previous work where each subtask is tackled in isolation

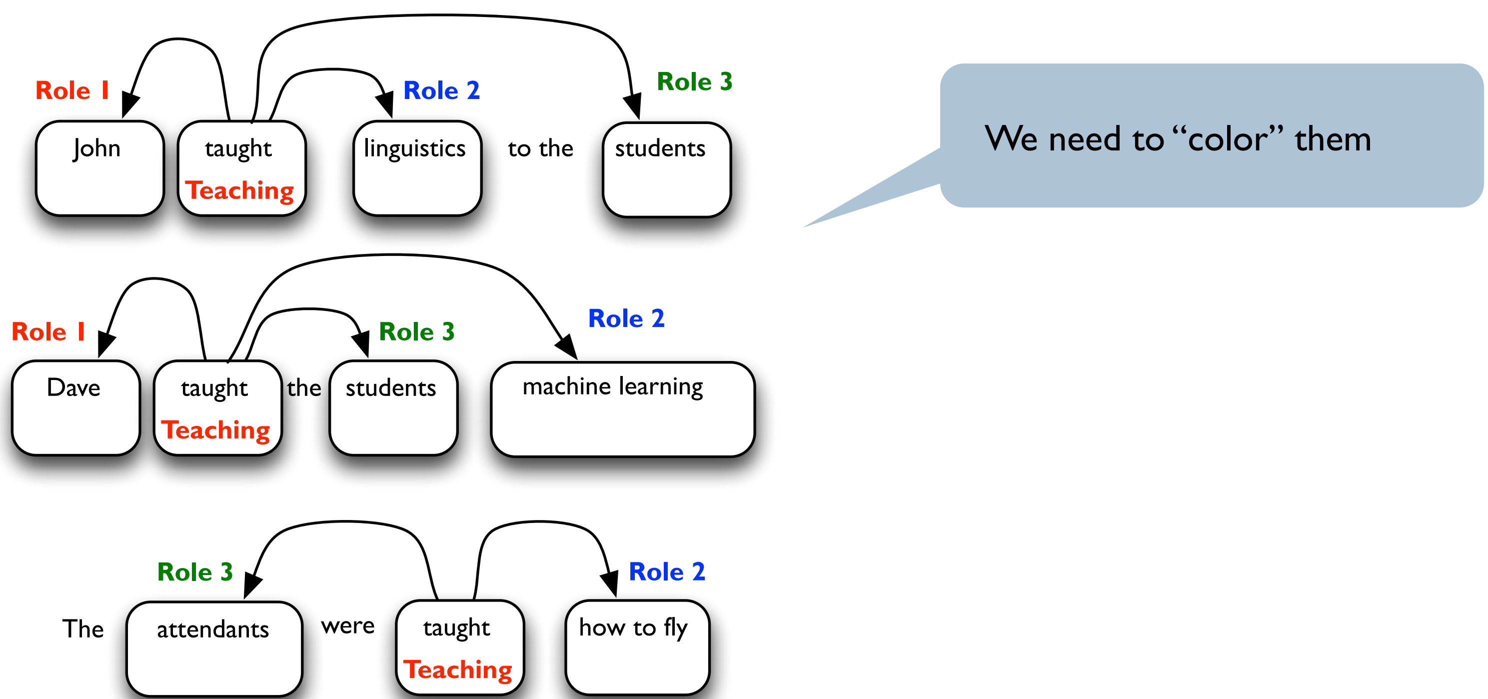
Induction of Semantic Roles: Definition

- Though after argument and semantic class identification and we know where arguments are, we do not know their semantic roles
- The step can be regarded as clustering of argument occurrences for a given semantic class



Induction of Semantic Roles: Definition

- Though after argument and semantic class identification and we know where arguments are, we do not know their semantic roles
- The step can be regarded as clustering of argument occurrences <u>for a given semantic class</u>

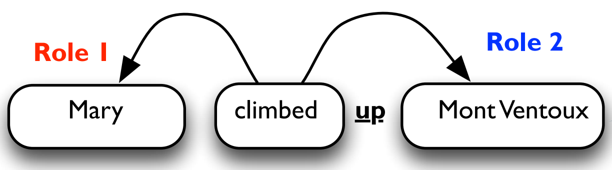


- The search space is huge – in realistic datasets frequents semantic classes appear tens of thousands times

Role Labeling as Clustering of Argument Keys

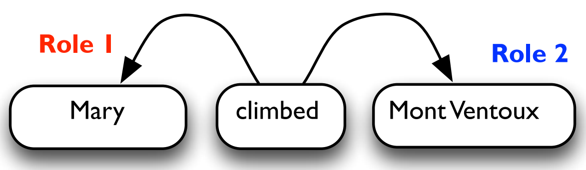
[Lang and Lapata, 2011b, Titov and Klementiev, 2011]

- Associate argument occurrences with syntactic signatures or <u>argument keys</u>
  - Will include simple syntactic cues such as verb voice and position relative to predicate



`ACTIVE:RIGHT:PMOD_up`

We assume the automatic syntactic analyses are available



`ACTIVE:RIGHT:OBJ`

Purity of around 90%

- Argument keys are designed to map to a single semantic role as much as possible (for an individual predicate)

All *occurrences* with the same key are automatically in the same cluster

Instead of clustering argument occurrences, the method clusters their argument keys

- Here, we would cluster `ACTIVE:RIGHT:OBJ` and `ACTIVE:RIGHT:PMOD_up` together

[Titov and Klementiev, 2012a]

A Bayesian model for role labeling

- Idea: propose a generative model for inducing argument clusters
  - clusters are of argument keys, not argument occurrences

- Learning signals:
  - Selection preferences

    i.e. distribution of argument fillers is sparse for every role

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

    John taught students math

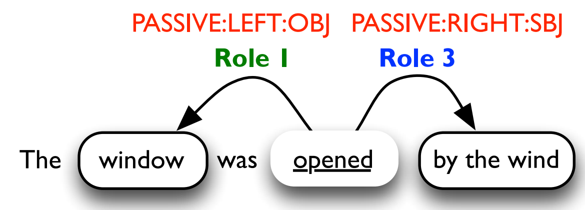
    GB-criterion

  - Syntax is predictive of roles

- How can we encode these signals in a generative story?

[Titov and Klementiev, 2012a]

A Bayesian model for role labeling



At least one argument

Draw first argument

Continue generation

Draw more arguments

Decide on arg key clustering

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

for each predicate $p = 1, 2, \ldots$:
$B_p \sim CRP(\alpha)$

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

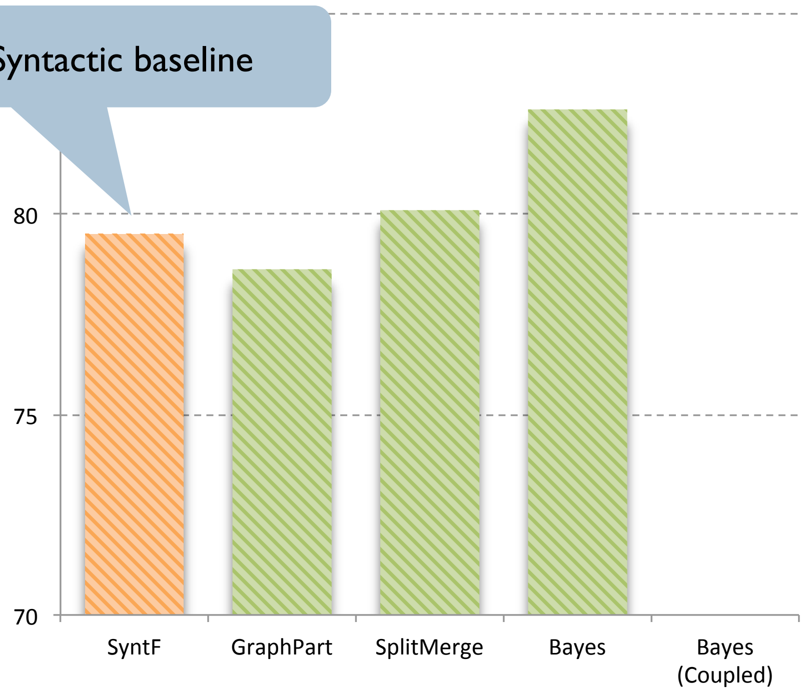
for each predicate $p = 1, 2, \ldots$:
for each role $r \in B_p$:
$\theta_{p,r} \sim DP(\beta, H^{(A)})$
$\psi_{p,r} \sim Beta(\eta_0, \eta_1)$

PropBank (CoNLL 08)

Clustering F1, Harmonic mean of purity and collocation

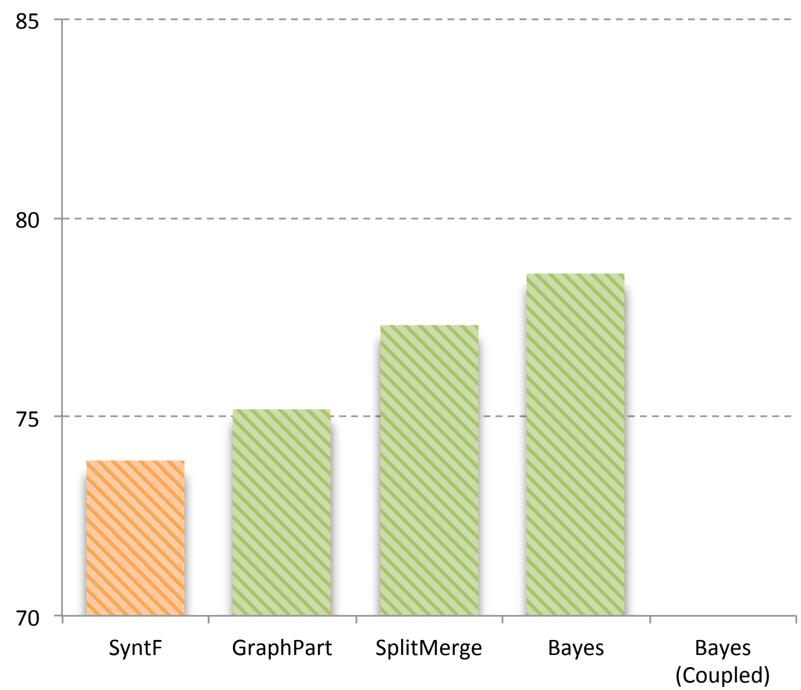
Gold syntax

Syntactic baseline



Previous approaches

Predicted syntax



A Bayesian model for role labeling

- The approaches we discussed induce roles for each predicate independently
- These clusterings define permissible *alternations*
- But many alternations are shared across verbs

or changes in the syntactic realizations of the argument structure of the verb

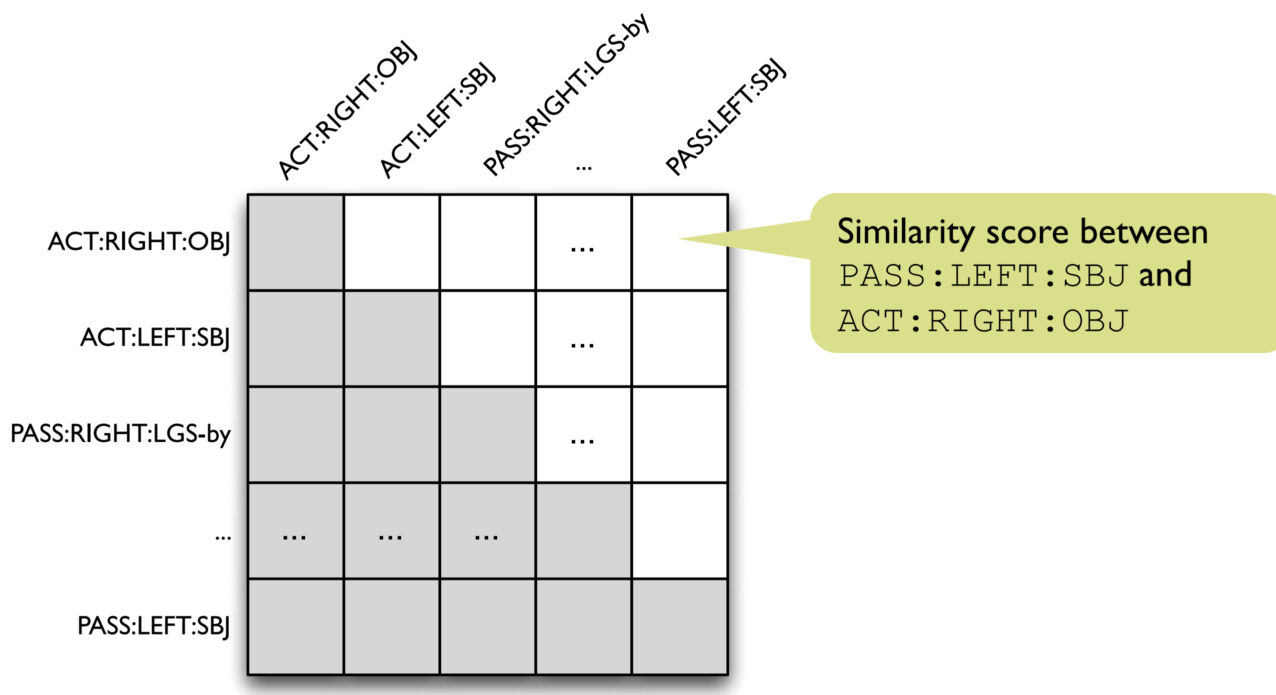
John gave the book to Mary	*vs*	John gave Mary the book
Mike threw the ball to me	*vs*	Mike threw me the ball

Dative alternation

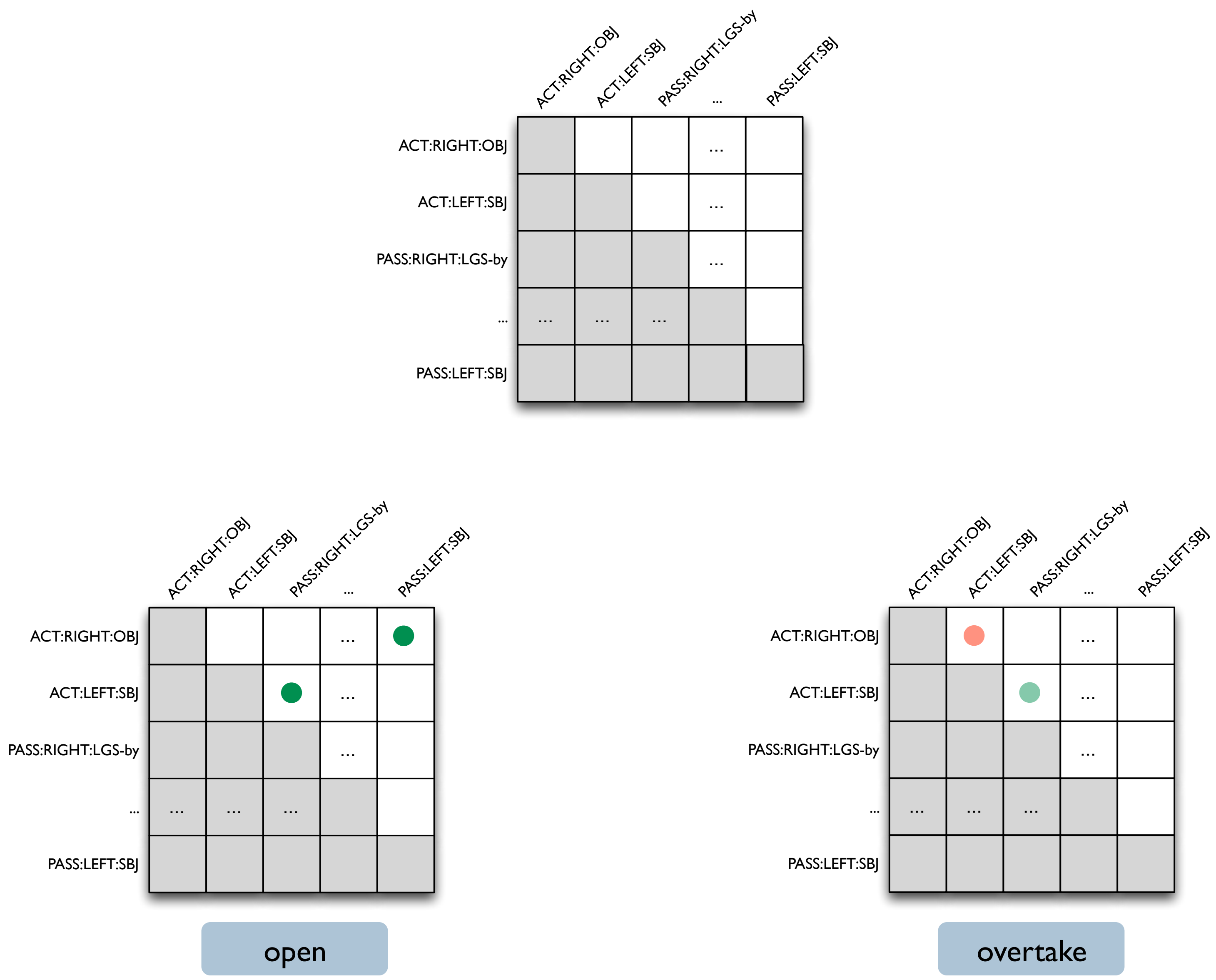
- Can we share this information across verbs?

A Bayesian model for role labeling

- Idea: keep track of how likely a pair of argument keys should be clustered
  - Define a similarity matrix (or similarity graph)

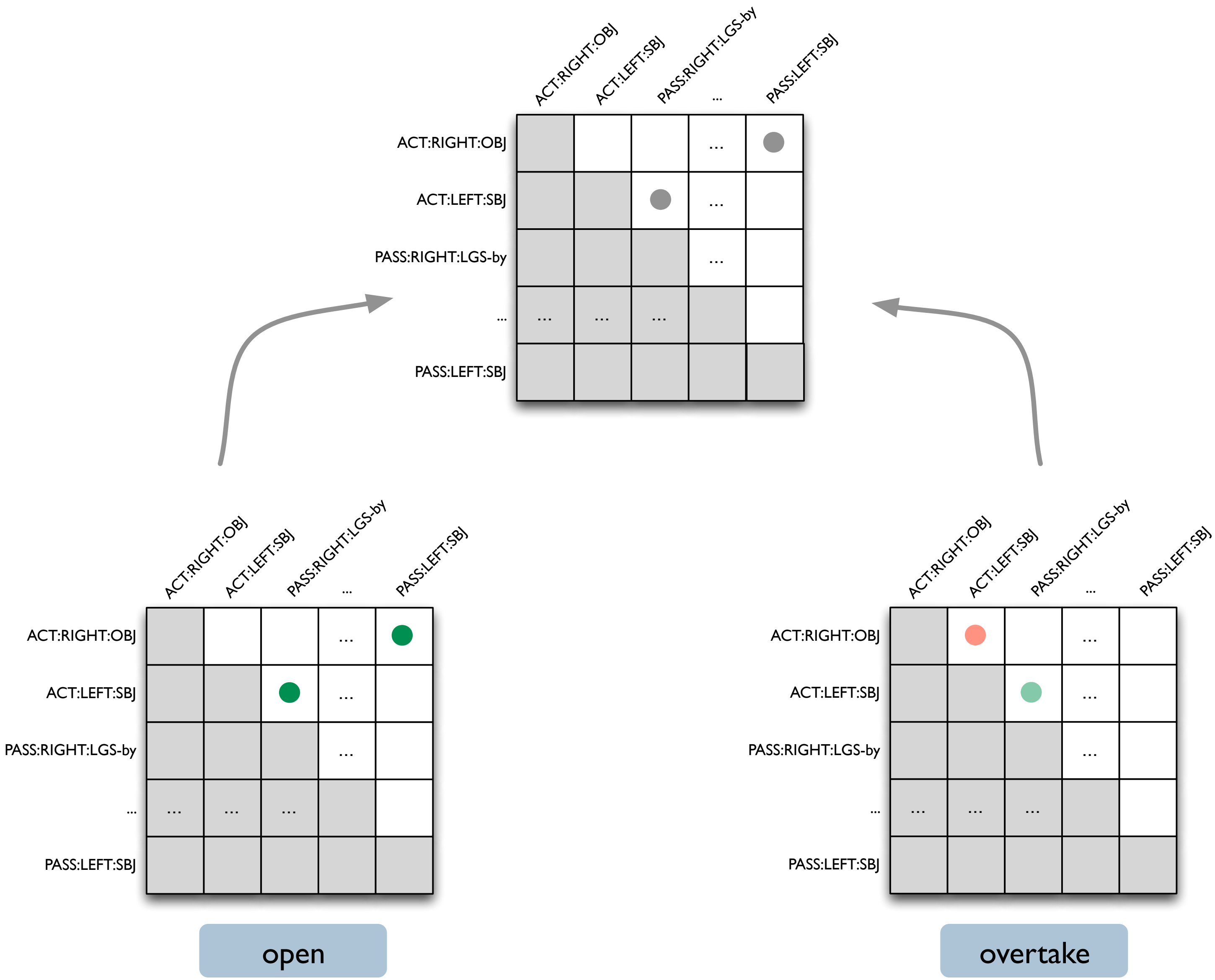


A Bayesian model for role labeling

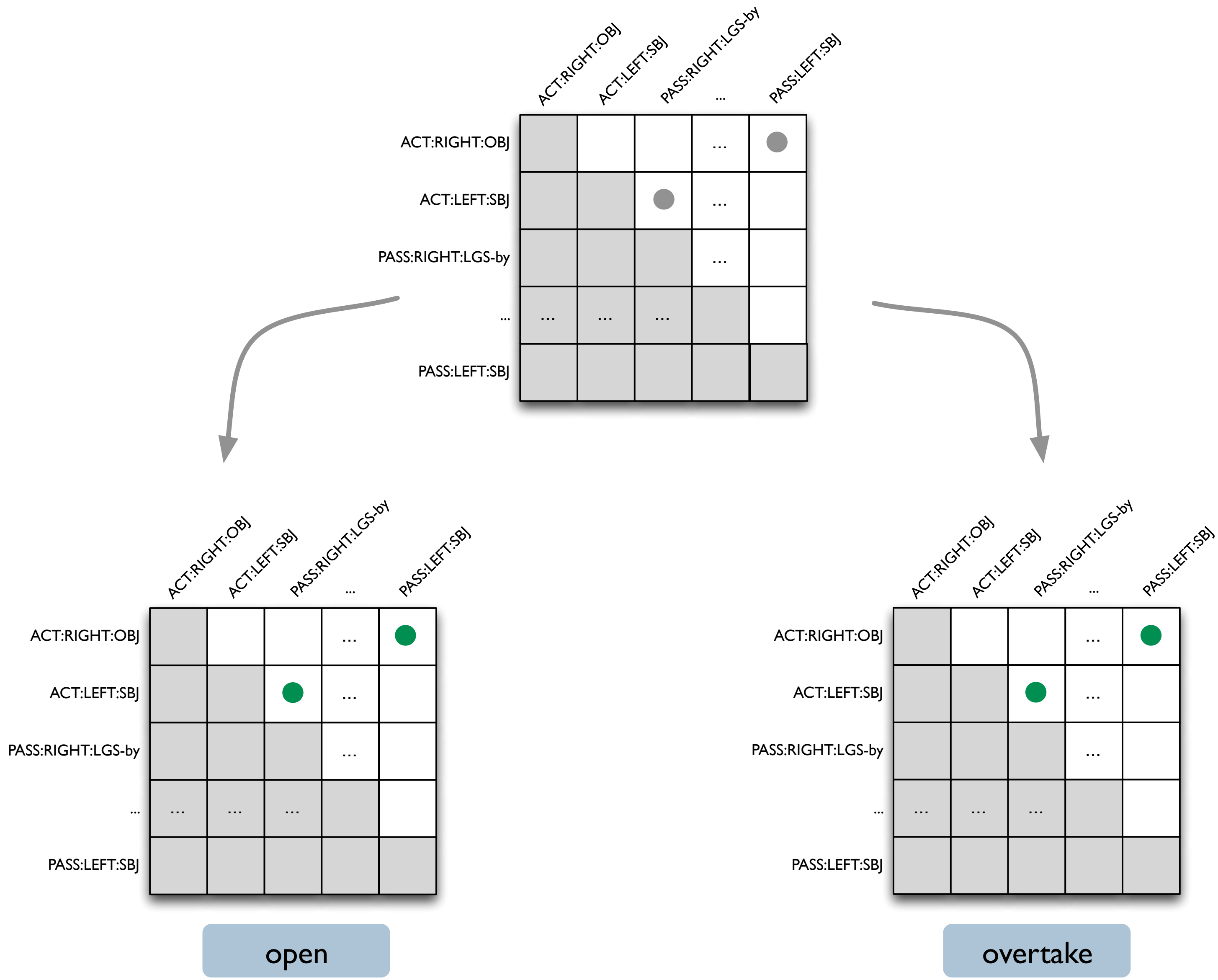


[Titov and Klementiev, 2012a]

A Bayesian model for role labeling



A Bayesian model for role labeling



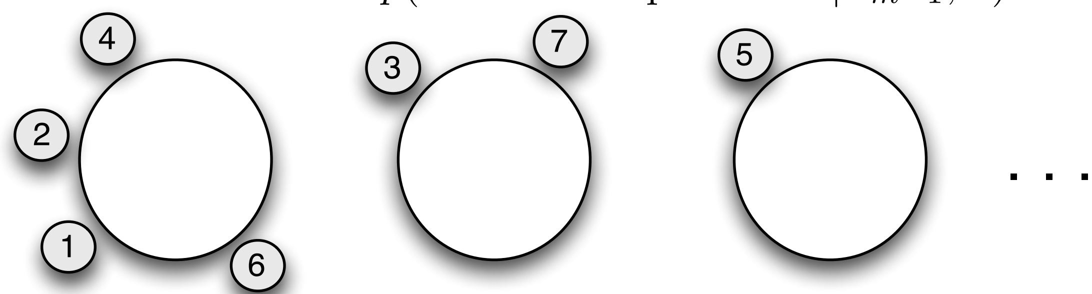
A formal way to encode this: dd-CRP

- 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

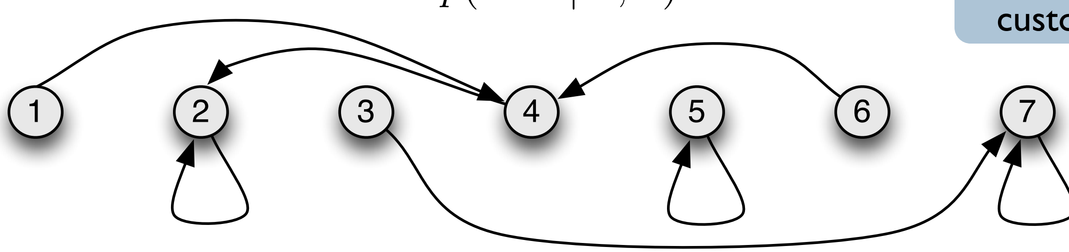


- An extension is distance-dependent CRP (dd-CRP):
  - m-th customer chooses a *customer* to sit with according to:

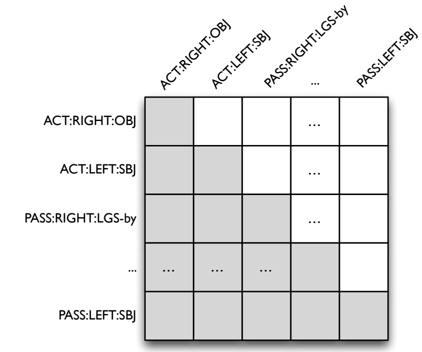
Entire similarity graph

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

Similarity between customers m and j



Sharing roles



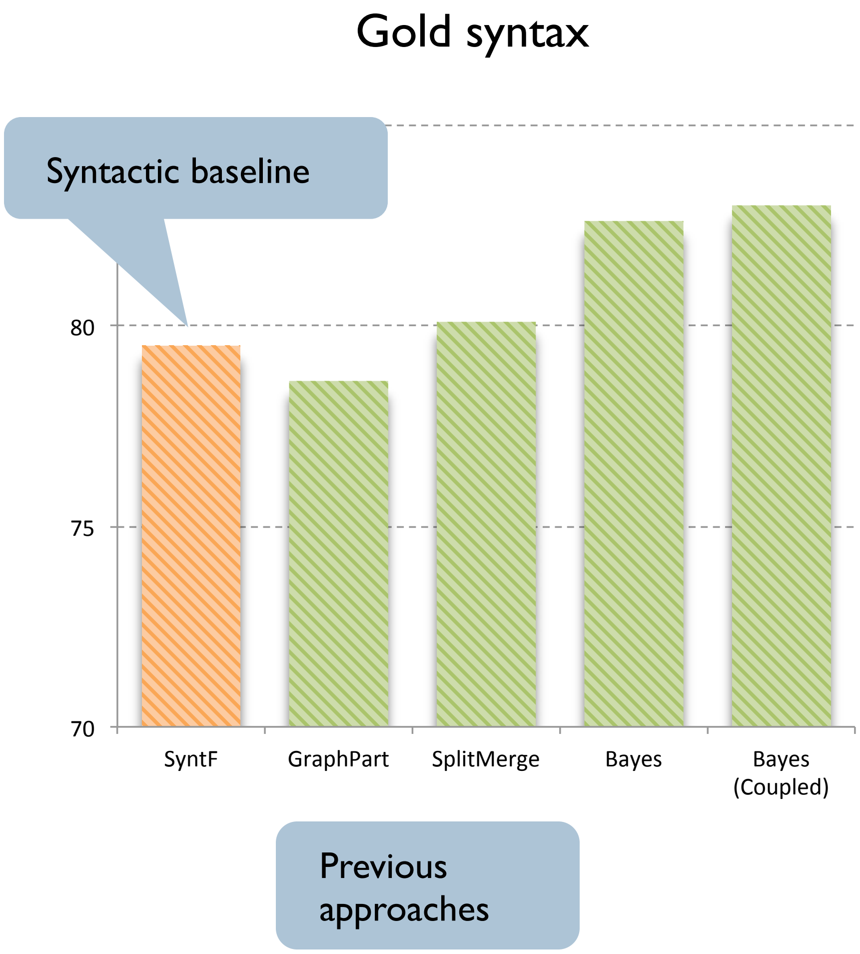
- Similarity graph *D* to couples distinct but similar clusterings of argument keys across predicates
  - Vertices are argument keys
  - Weights are similarity scores for each pair of argument keys

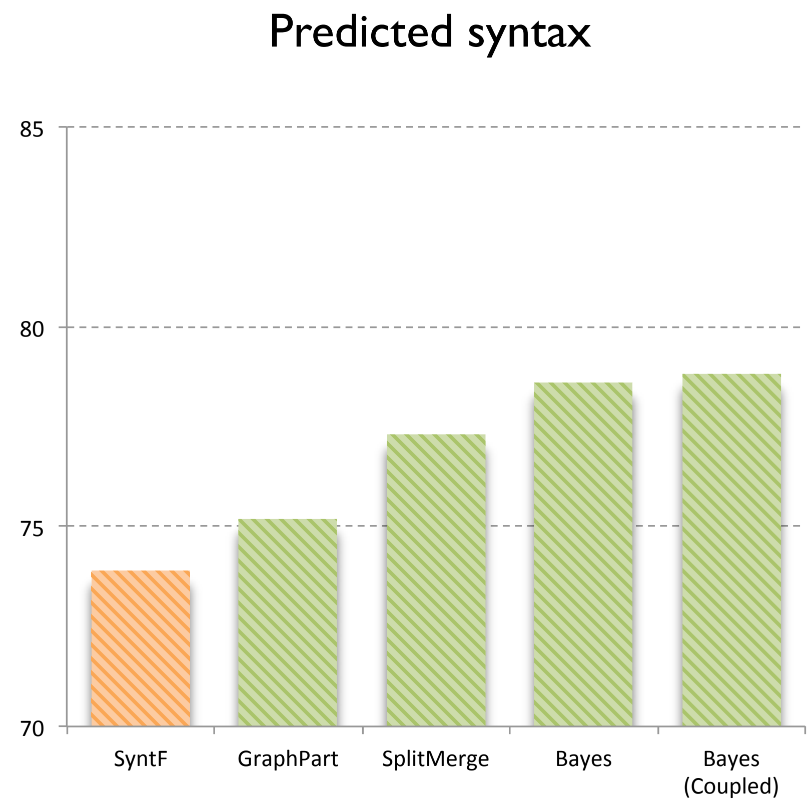
- We treat *D* as a latent random variable drawn from a prior over weighted graphs
  - First drawn from a prior
  - Used to generate each of the clusterings for every predicate

- 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

[Titov and Klementiev, 2012a]

PropBank (CoNLL 08)





Qualitative

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

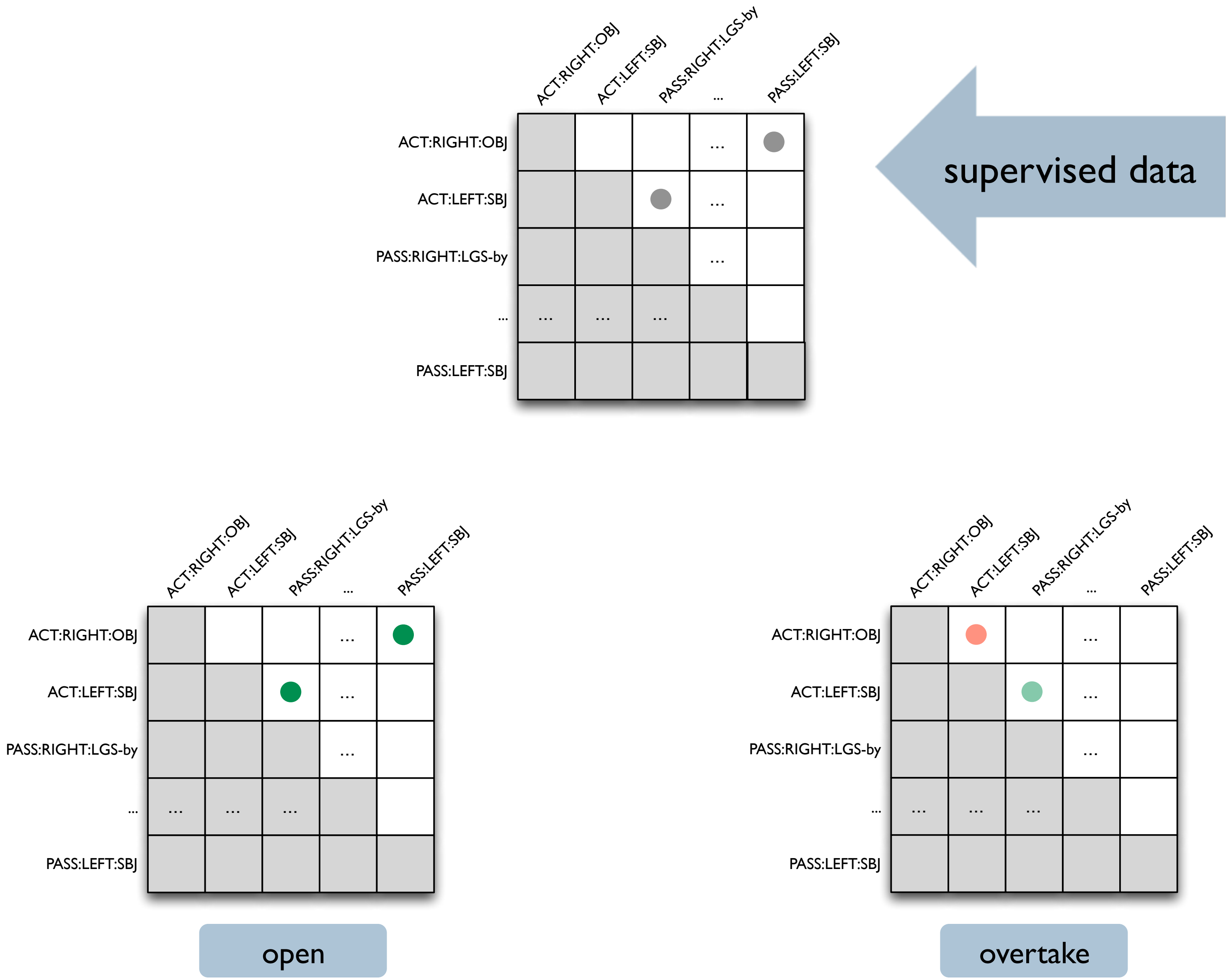
  Martha carved a doll for the baby

  Martha carved the baby a doll
- Dativization

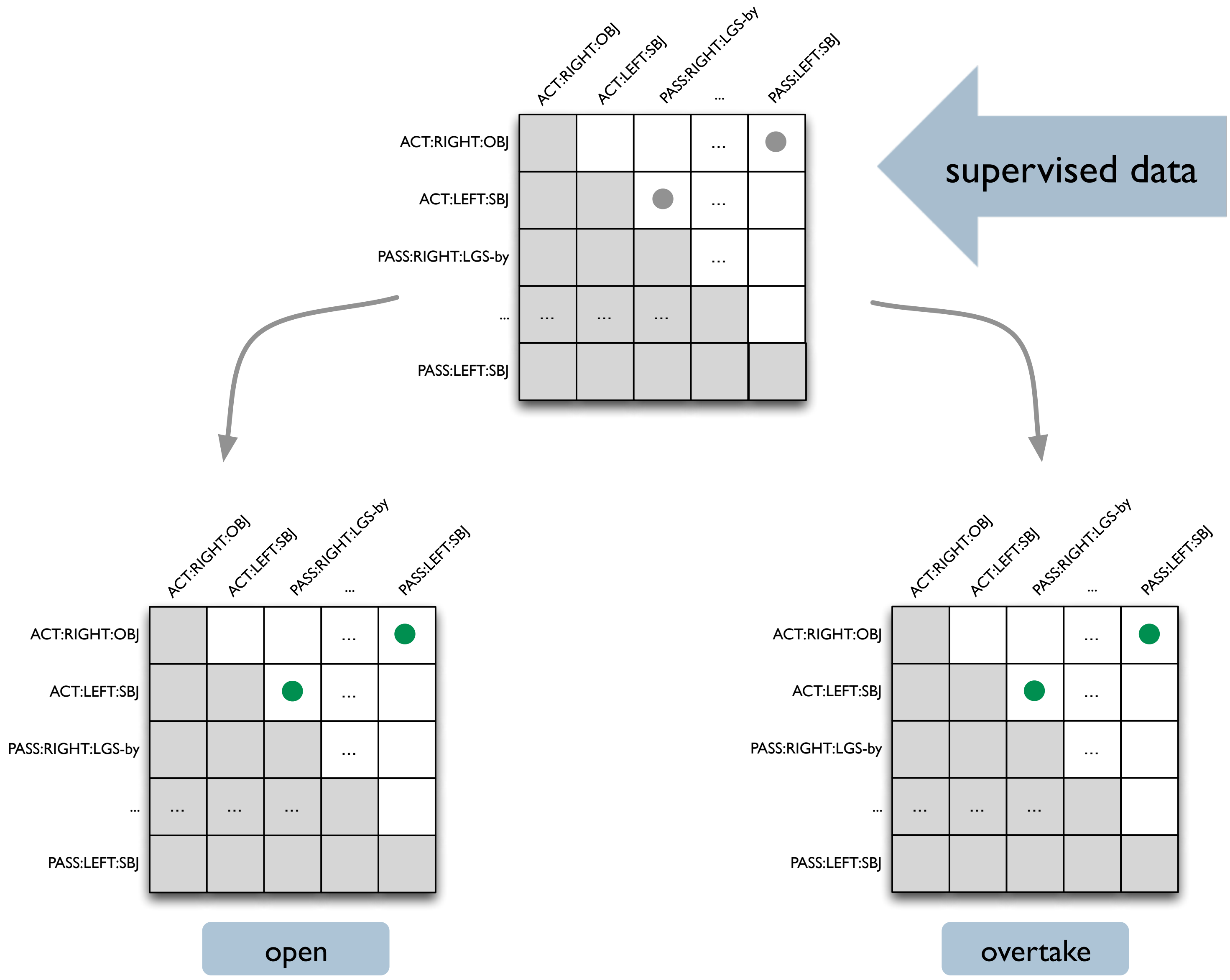
  I gave the book to Mary

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

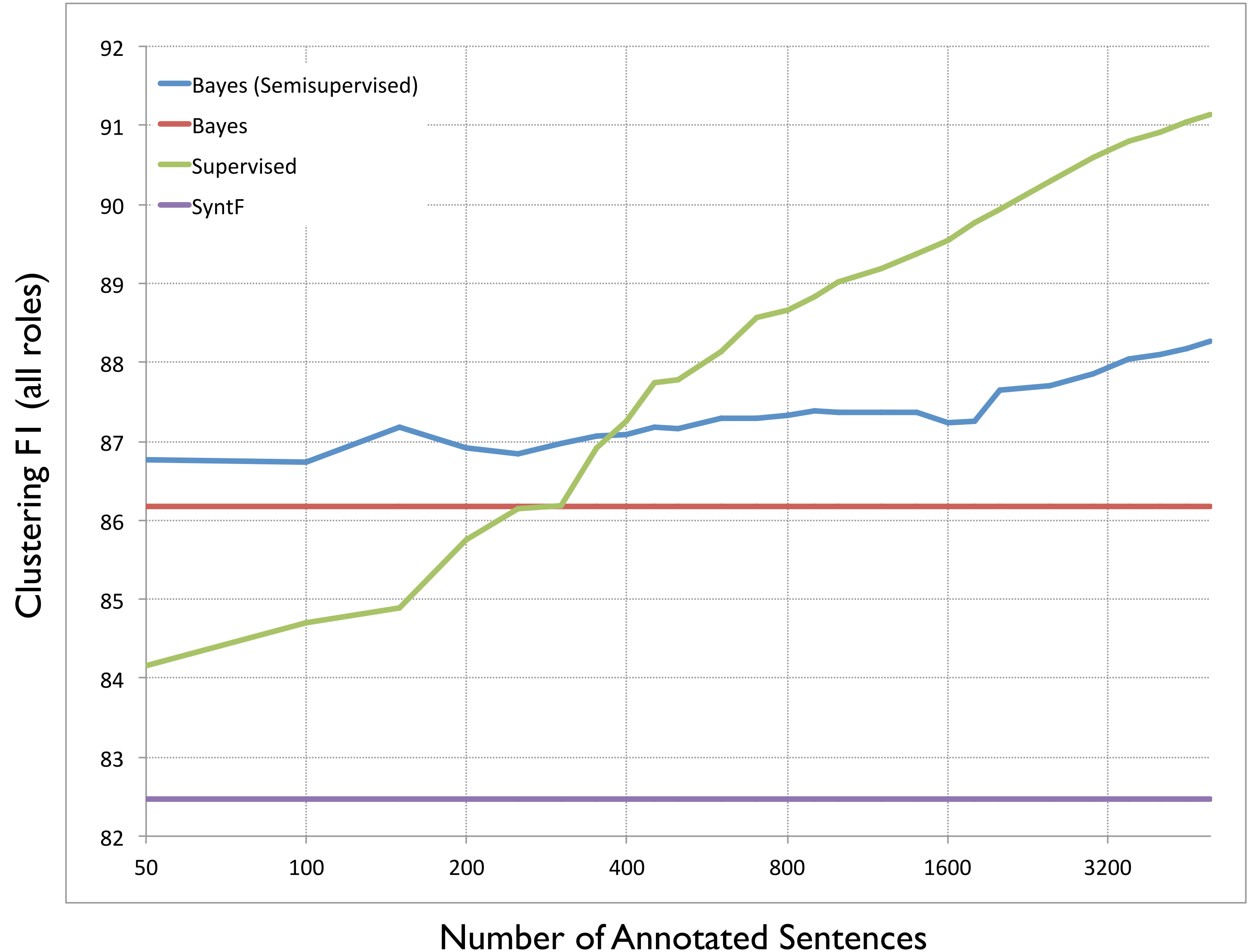
A Bayesian model for role labeling



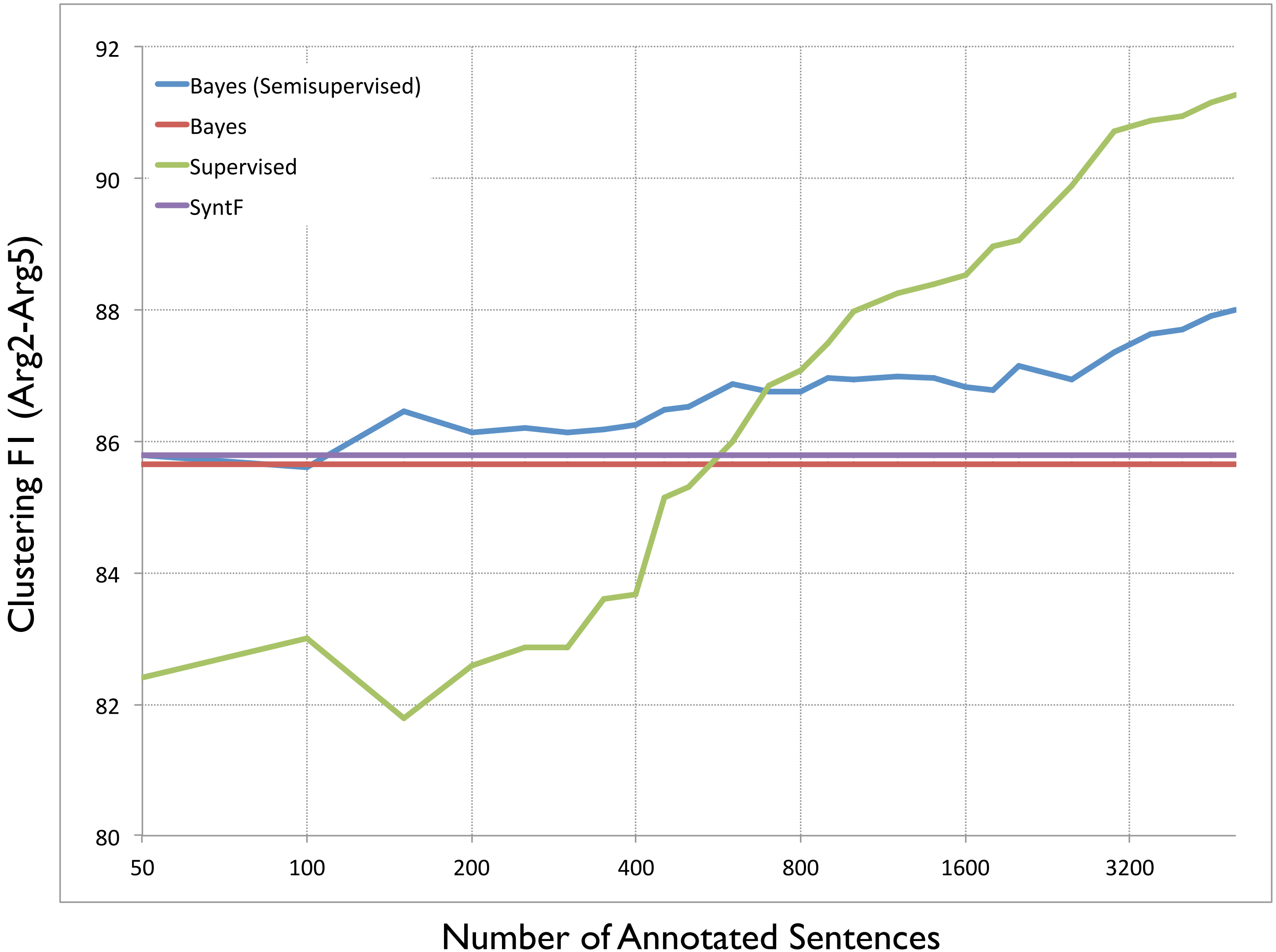
A Bayesian model for role labeling



PropBank (CoNLL 09)



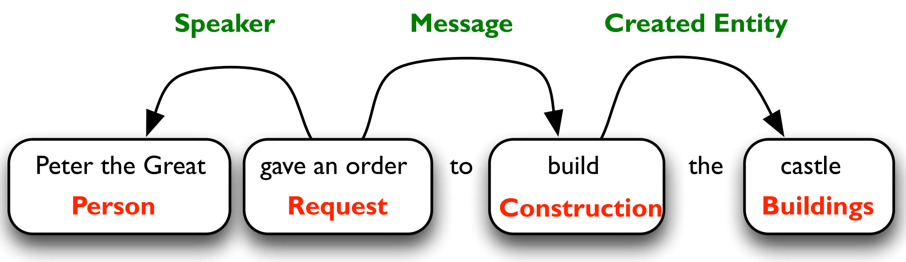
PropBank (CoNLL 09)



Outline

- Induction of events and their participants
  - unsupervised models of semantic roles
  - joint induction of frames and roles
  - cross-lingual extension and comparison with projection and transfer
- Induction of semantic representations of words and phrases
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, lexicon induction)

Induction of frames / semantic classes

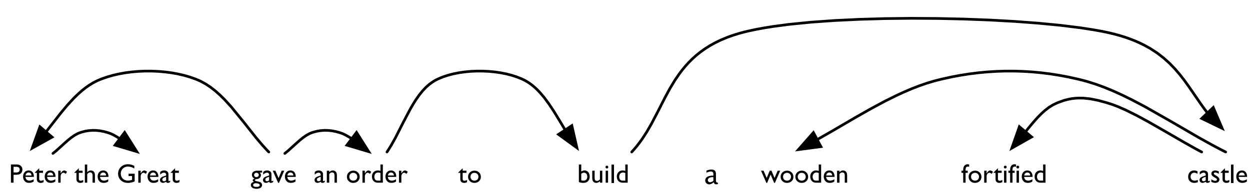


- Induction of frames and induction of argument clusters – the same task
  - We will refer to both of them as semantic classes
- Induction of semantic classes involves:
  - Clustering of lexemes with similar meaning
    - *break*, *bust*, *destroy* should be clustered together
  - Detection of multi-word expression, i.e. expressions which are not (sufficiently) compositional
    - these includes idiomatic expressions, terminology, proper nouns, …
    - E.g., *hold a victory over*, *red herring*

Later, they can be clustered with atomic ones.
E.g., win + *held a victory over*

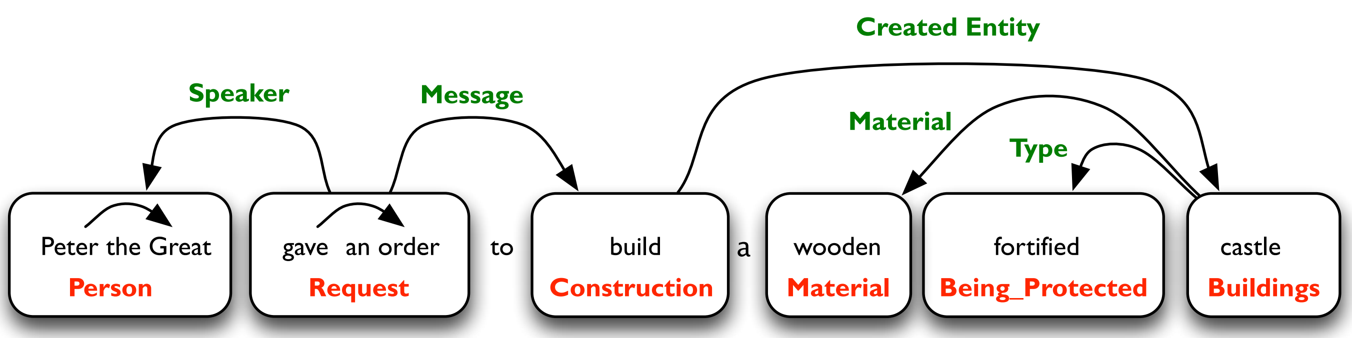
Generalization of the role induction model

- The model can be generalized for joint induction of predicate-argument structure of an entire sentence
  - start with a (transformed) syntactic dependency graph (~ argument identification)



Generalization of the role induction model

- The model can be generalized for joint induction of predicate-argument structure of an entire sentence
  - start with a (transformed) syntactic dependency graph (~ argument identification)
  - predict decomposition and labeling of its parts
    - label on nodes are frames (or *semantic classes* of arguments)
    - labels on edges are roles (frame elements)



The Joint Model

Draw semantic class for root

$$
\begin{aligned}
&for\ each\ sentence: \\
&\quad c_{root} \sim \theta_{root} \\
&\quad \textbf{GenSemClass}(c_{root})
\end{aligned}
$$

The Joint Model

Draw semantic class for root

$for\ each\ sentence:$

$c_{root} \sim \theta_{root}$

$\mathbf{GenSemClass}(c_{root})$

$\mathbf{GenSemClass}(c)$

$s \sim \phi_c$

$for\ each\ role\ t = 1, \ldots, T:$

$if\ [n \sim \psi_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$while\ [n \sim \psi^{+}_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

Request

The Joint Model

Draw semantic class for root

$for\ each\ sentence:$

$c_{root} \sim \theta_{root}$

$\mathbf{GenSemClass}(c_{root})$

Draw synt/lex realization

$\mathbf{GenSemClass}(c)$

$s \sim \phi_c$

$for\ each\ role\ t = 1, \ldots, T:$

$if\ [n \sim \psi_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$while\ [n \sim \psi^{+}_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

{ We use hierarchical Dirichlet processes to represent distributions over tree fragments }

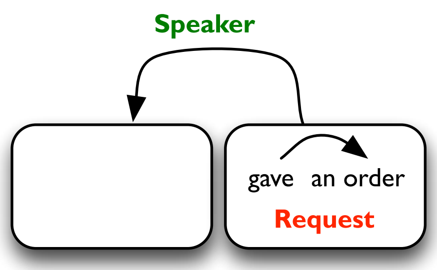
The Joint Model

Draw semantic class for root

$for\ each\ sentence:$

$c_{root} \sim \theta_{root}$

$\mathbf{GenSemClass}(c_{root})$



$\mathbf{GenSemClass}(c)$

Draw synt/lex realization

At least one argument

$s \sim \phi_c$

$for\ each\ role\ t = 1, \dots, T:$

$if\ [n \sim \psi_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$while\ [n \sim \psi^{+}_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

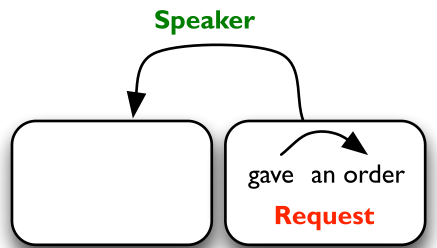
The Joint Model

Draw semantic class for root

for each sentence :

$c_{root} \sim \theta_{root}$

$\mathbf{GenSemClass}(c_{root})$



GenSemClass(c)

Draw synt/lex realization

$s \sim \phi_c$

At least one argument

for each role $t = 1, \ldots, T$:

$if\ [n \sim \psi_{c,t}] = 1$:

Draw first argument

$\mathbf{GenArgument}(c,t)$

$while\ [n \sim \psi^{+}_{c,t}] = 1$:

$\mathbf{GenArgument}(c,t)$

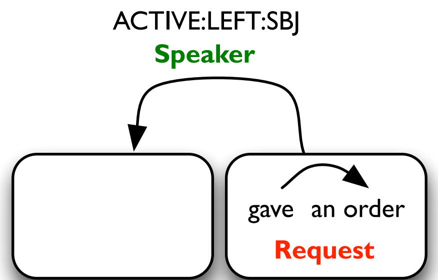
GenArgument(c,t)

$a_{c,t} \sim \phi_{c,t}$

$c'_{c,t} \sim \theta_{c,t}$

$\mathbf{GenSemClass}(c'_{c,t})$

The Joint Model



Draw semantic class for root

$$\textit{for each sentence}:$$
$$c_{root} \sim \theta_{root}$$
$$\mathbf{GenSemClass}(c_{root})$$

Draw synt/lex realization

At least one argument

Draw first argument

$$\mathbf{GenSemClass}(c)$$

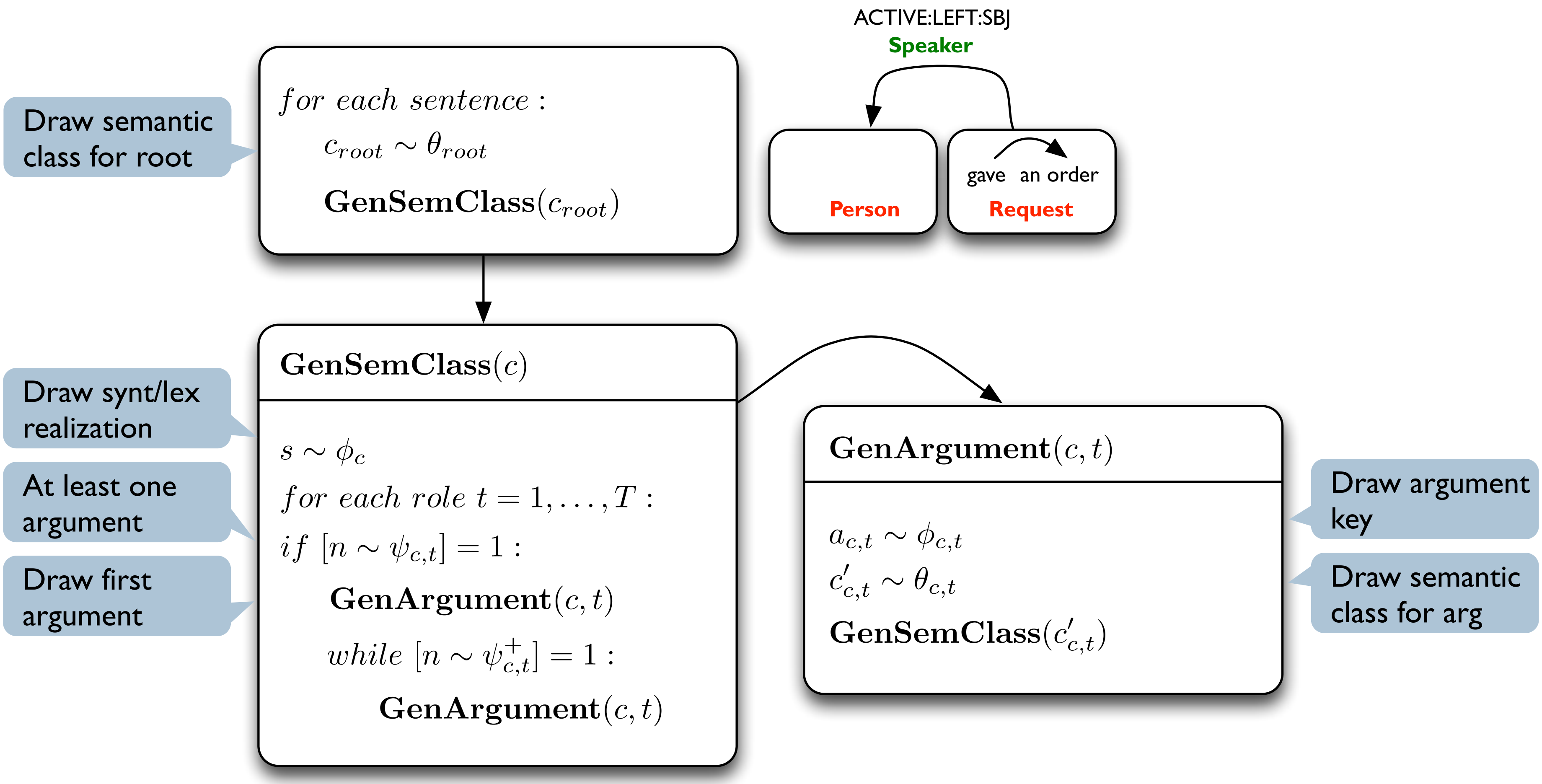
$$s \sim \phi_c$$
$$\textit{for each role } t = 1, \ldots, T:$$
$$\textit{if } [n \sim \psi_{c,t}] = 1:$$
$$\mathbf{GenArgument}(c, t)$$
$$\textit{while } [n \sim \psi^{+}_{c,t}] = 1:$$
$$\mathbf{GenArgument}(c, t)$$

$$\mathbf{GenArgument}(c, t)$$

$$a_{c,t} \sim \phi_{c,t}$$
$$c'_{c,t} \sim \theta_{c,t}$$
$$\mathbf{GenSemClass}(c'_{c,t})$$

Draw argument key

The Joint Model



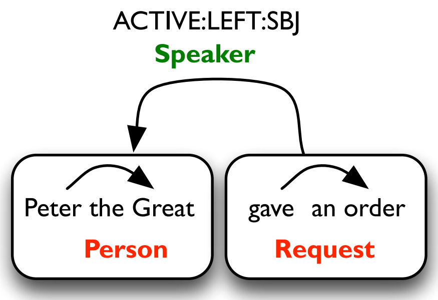
The Joint Model

Draw semantic class for root

$for\ each\ sentence:$

$c_{root} \sim \theta_{root}$

$\mathbf{GenSemClass}(c_{root})$



Draw synt/lex realization

At least one argument

Draw first argument

$\mathbf{GenSemClass}(c)$

$s \sim \phi_c$

$for\ each\ role\ t = 1, \ldots, T:$

$if\ [n \sim \psi_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$while\ [n \sim \psi^{+}_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$\mathbf{GenArgument}(c, t)$

$a_{c,t} \sim \phi_{c,t}$

$c'_{c,t} \sim \theta_{c,t}$

$\mathbf{GenSemClass}(c'_{c,t})$

Draw argument key

Draw semantic class for arg

Recurse

The Joint Model

Draw semantic class for root

$for\ each\ sentence:$

$c_{root} \sim \theta_{root}$

$\mathbf{GenSemClass}(c_{root})$

ACTIVE:LEFT:SBJ
Speaker

Peter the Great
Person

gave an order
Request

Draw synt/lex realization

At least one argument

Draw first argument

Continue generation

$\mathbf{GenSemClass}(c)$

$s \sim \phi_c$

$for\ each\ role\ t = 1, \ldots, T:$

$if\ [n \sim \psi_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$while\ [n \sim \psi^{+}_{c,t}] = 1:$

$\mathbf{GenArgument}(c, t)$

$\mathbf{GenArgument}(c, t)$

$a_{c,t} \sim \phi_{c,t}$

$c'_{c,t} \sim \theta_{c,t}$

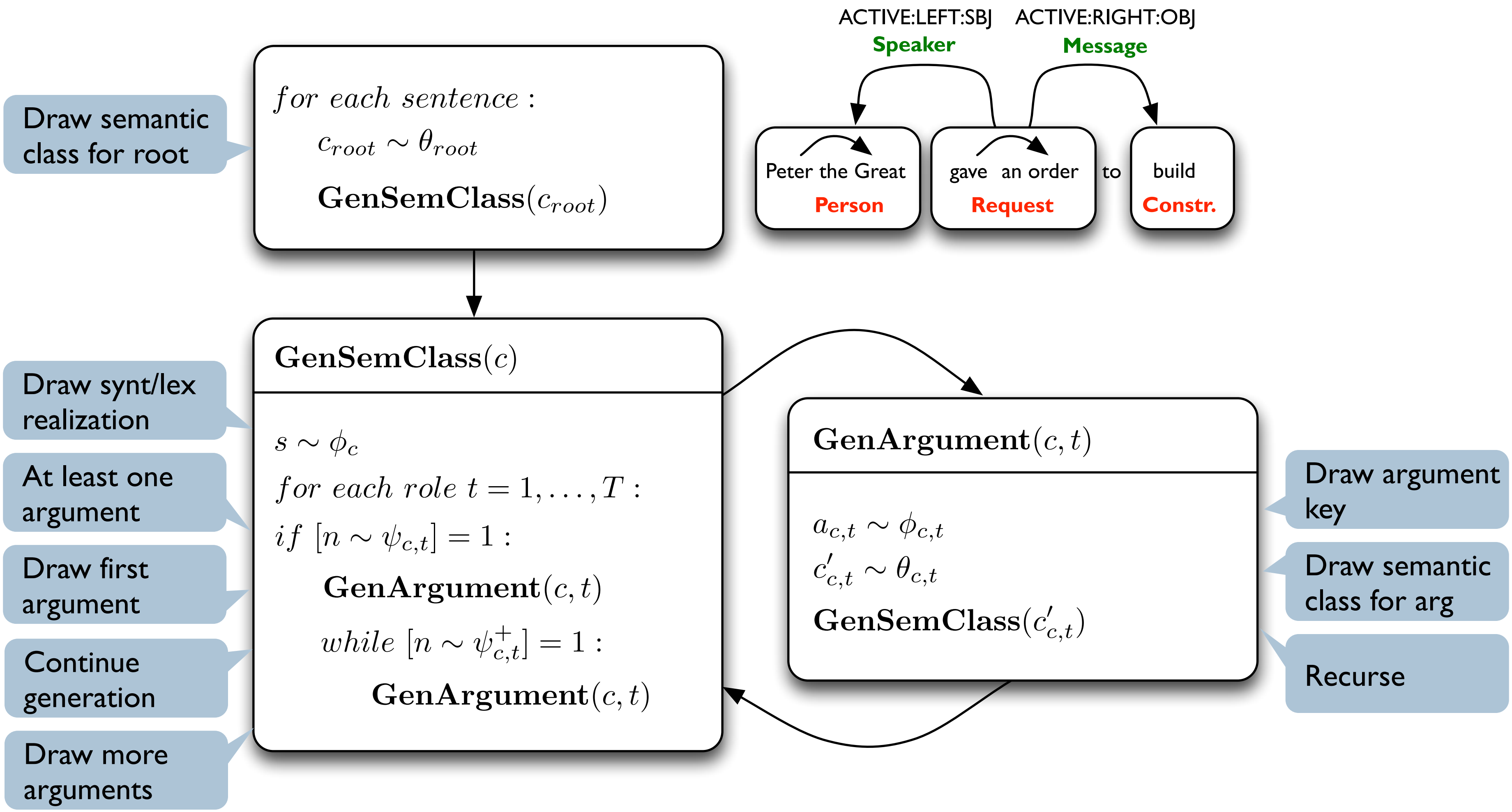
$\mathbf{GenSemClass}(c'_{c,t})$

Draw argument key

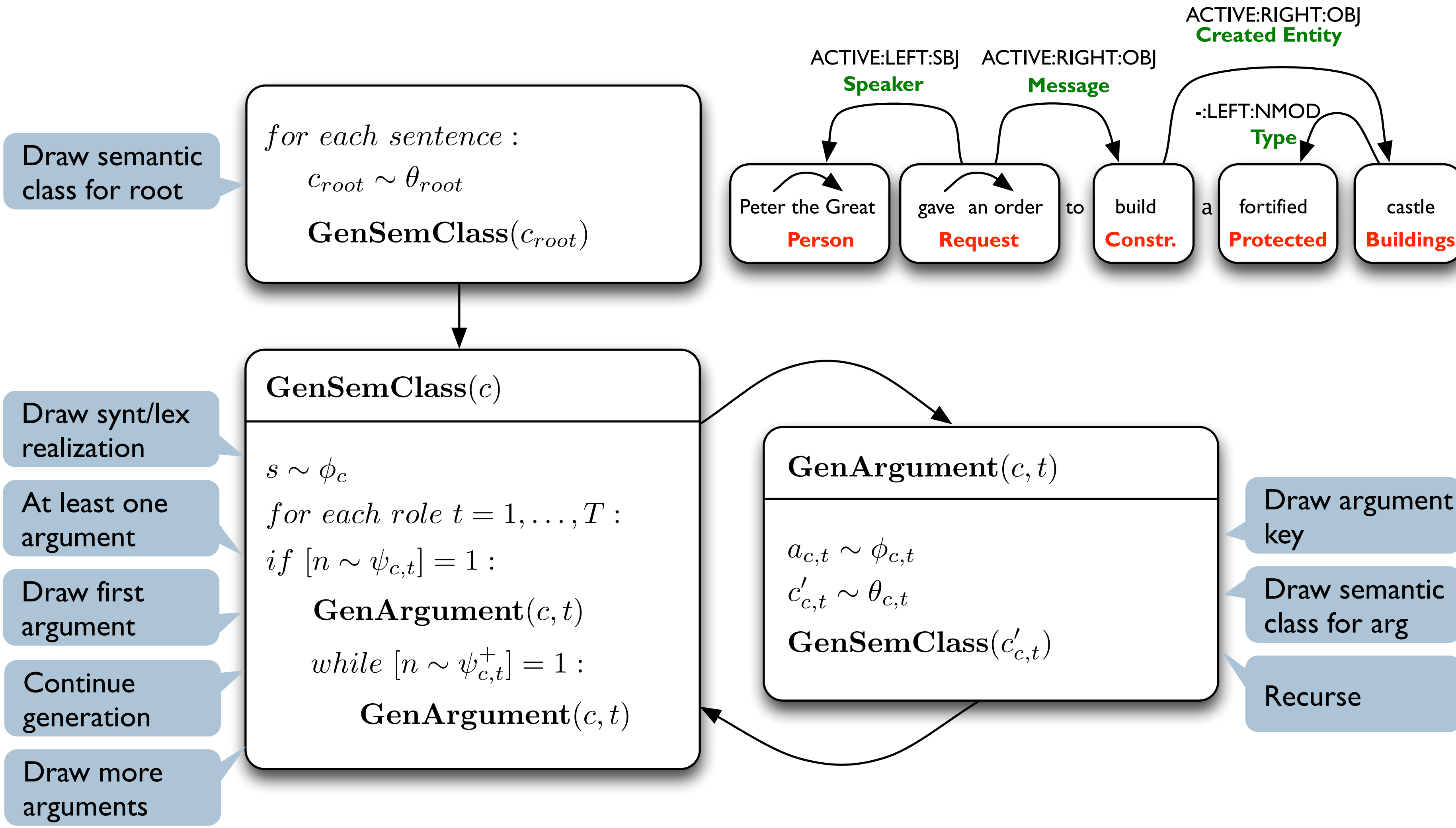
Draw semantic class for arg

Recurse

The Joint Model



The Joint Model



Inference

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

- Inference is challenging as the search space is huge
- We use a Metropolis-Hastings split-merge sampler with the following types of moves ('relabelings')
  - Role-Syntax alignment
    - Choose a new clustering of argument keys for a frame
  - Split – Merge
    - Merge 2 semantic classes together or split one class in two
  - Compose-Decompose
    - Compose fragments of syntactic tree to form a new realization or split a fragment

break + bust

held + a victory = held a victory

The similarity graph is also periodically updated

Application-based Evaluation

Question Answering about knowledge in a corpus of biomedical abstracts

- Dataset: 1999 biomedical abstracts from the Genia corpus (Kim et al, 2003)
- Examples of induced semantic classes:

Class	Variations
1	motif, sequence, regulatory element, response element, element, dna sequence
2	donor, individual, subject
3	important, essential, critical
4	dose, concentration
5	activation, transcriptional activation, transactivation
6	b cell, t lymphocyte, thymocyte, b lymphocyte, t cell, t-cell line, human lymphocyte, t-lymphocyte
7	indicate, reveal, document, suggest, demonstrate
8	augment, abolish, inhibit, convert, cause, abrogate, modulate, block, decrease, reduce, diminish, suppress, up-regulate, impair, reverse, enhance
9	confirm, assess, examine, study, evaluate, test, resolve, determine, investigate
10	nf-kappab, nf-kappa b, nfkappab, nf-kb

Blood cells

Roughly "cause change position on a scale" frame

[Poon and Domingos, 2008]

Application-based Evaluation

Question Answering about knowledge in a corpus of biomedical abstracts

- Example questions and answers:

Question: What does cyclosporin A suppress?

Answer: expression of EGR-2

Sentence: As with EGR-3 , expression of EGR-2 was blocked by cyclosporin A .

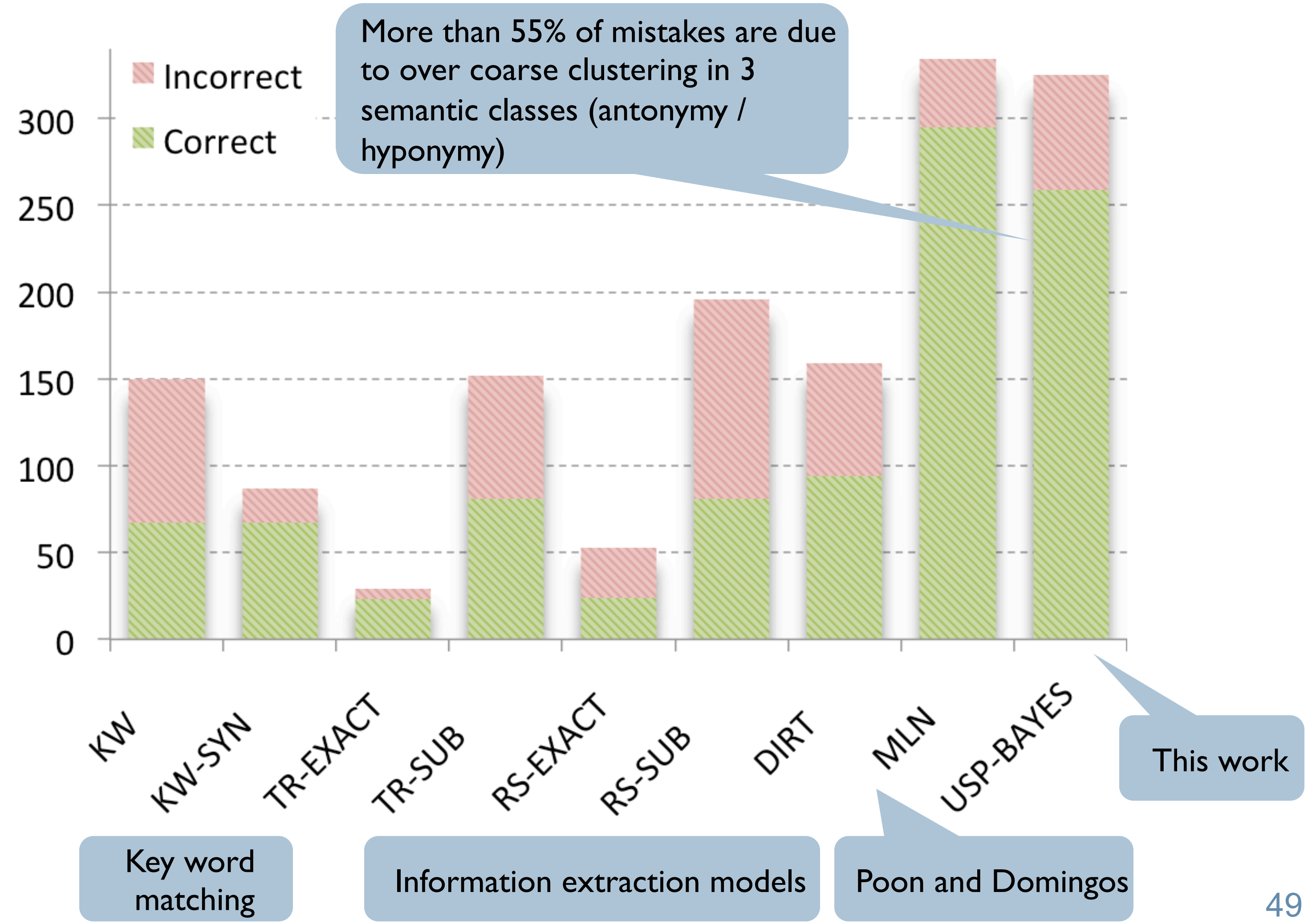
Question: What inhibits tnf-alpha?

Answer: IL -10

Sentence: Our previous studies in human monocytes have demonstrated that interleukin (IL) -10 inhibits lipopolysaccharide (LPS) -stimulated production of inflammatory cytokines , IL-1 beta , IL-6 , IL-8 , and tumor necrosis factor alpha by blocking gene transcription .

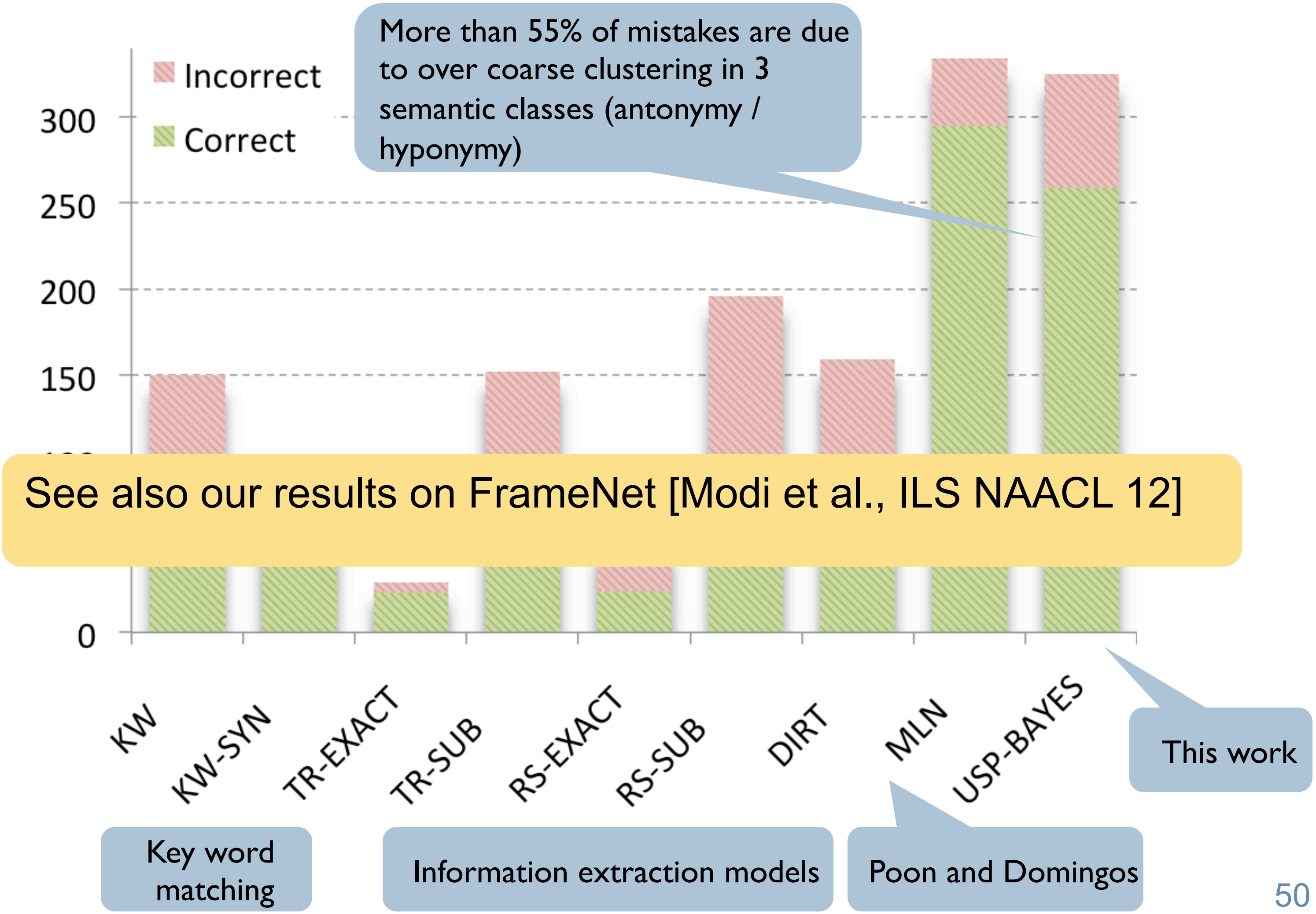
Application-based Evaluation

Question Answering about knowledge in a corpus of biomedical abstracts



Application-based Evaluation

Question Answering about knowledge in a corpus of biomedical abstracts

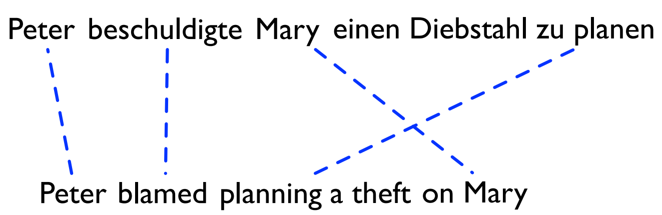


Outline

- Induction of events and their participants
  - unsupervised models of semantic roles
  - joint induction of frames and roles
  - cross-lingual extension and comparison with projection and transfer
- Induction of semantic representations of words and phrases
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, lexicon induction)

Crosslingual Induction of Semantic Roles

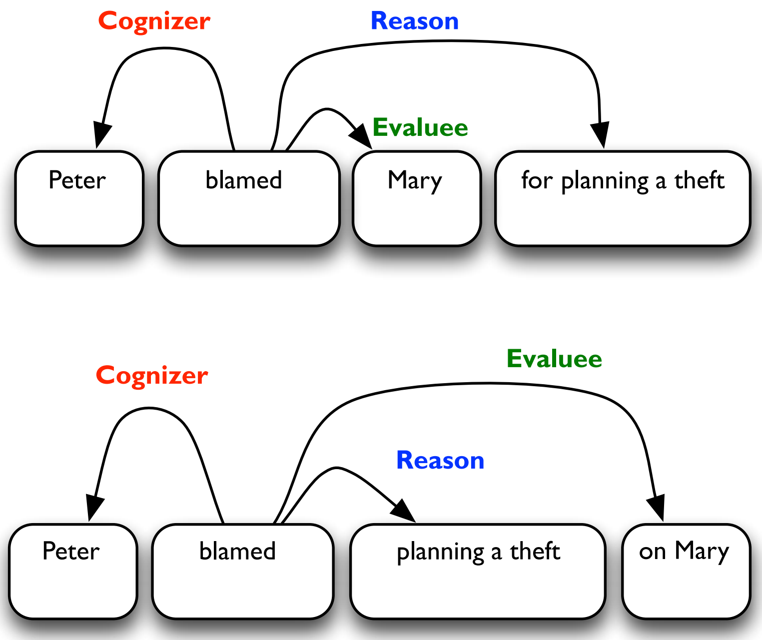
- We have additional multilingual resources: texts translated in multiple languages (parallel data)
  - Parliament proceedings, books, etc.
  - Can use standard machine translation techniques to induce word alignments



- We use aligned data and induce semantics jointly in multiple languages
  - Only during learning, we apply them to monolingual sentences

Crosslingual Induction of Semantic Roles

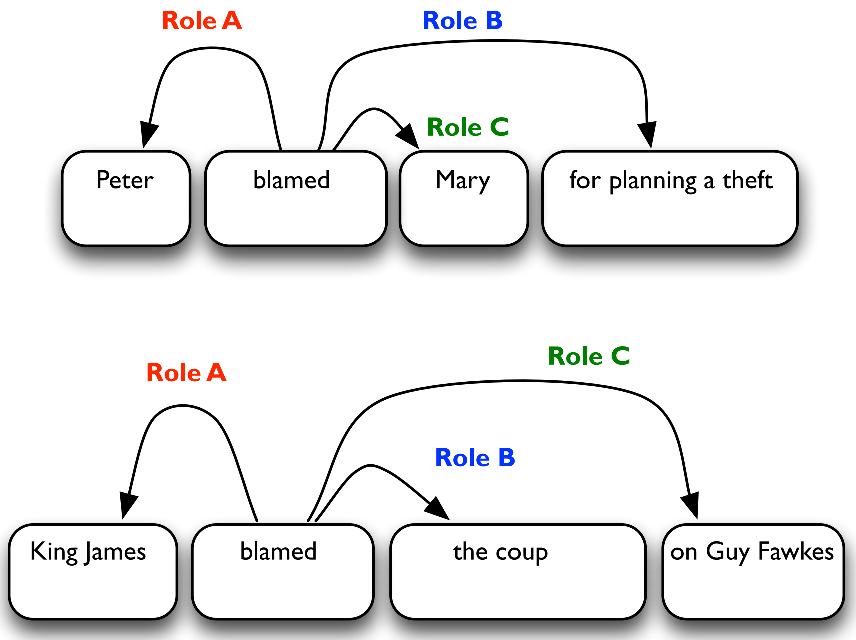
- Consider an example *blame* alternation



- Learning the corresponding linking is not trivial
  - selectional preferences for all roles are not very restrictive
  - selectional restrictions for Cognizer and Evaluee are overlapping

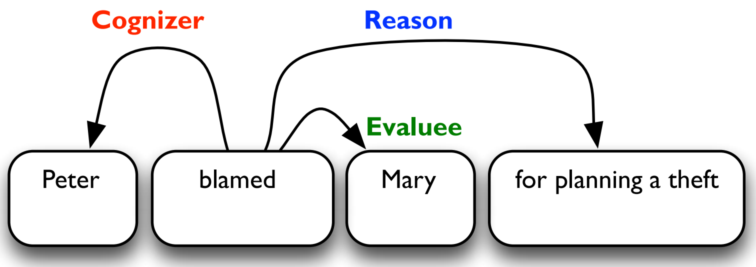
Crosslingual Induction of Semantic Roles

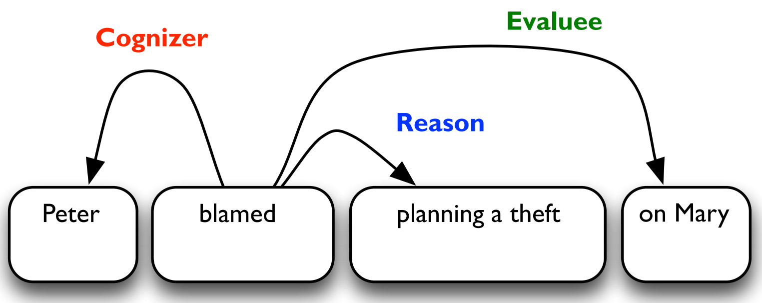
- Consider an example *blame* alternation



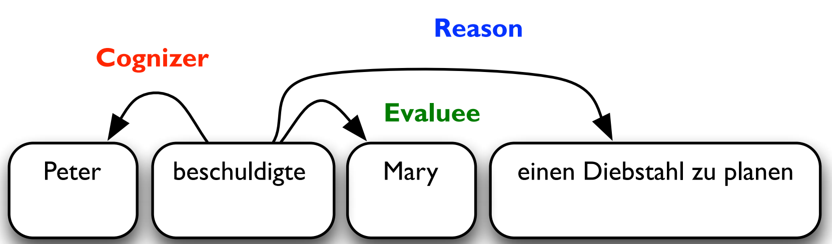
- Learning the corresponding linking is not trivial
  - selectional preferences for all roles are not very restrictive
  - selectional restrictions for Cognizer and Evaluee are overlapping

Crosslingual Induction of Semantic Roles



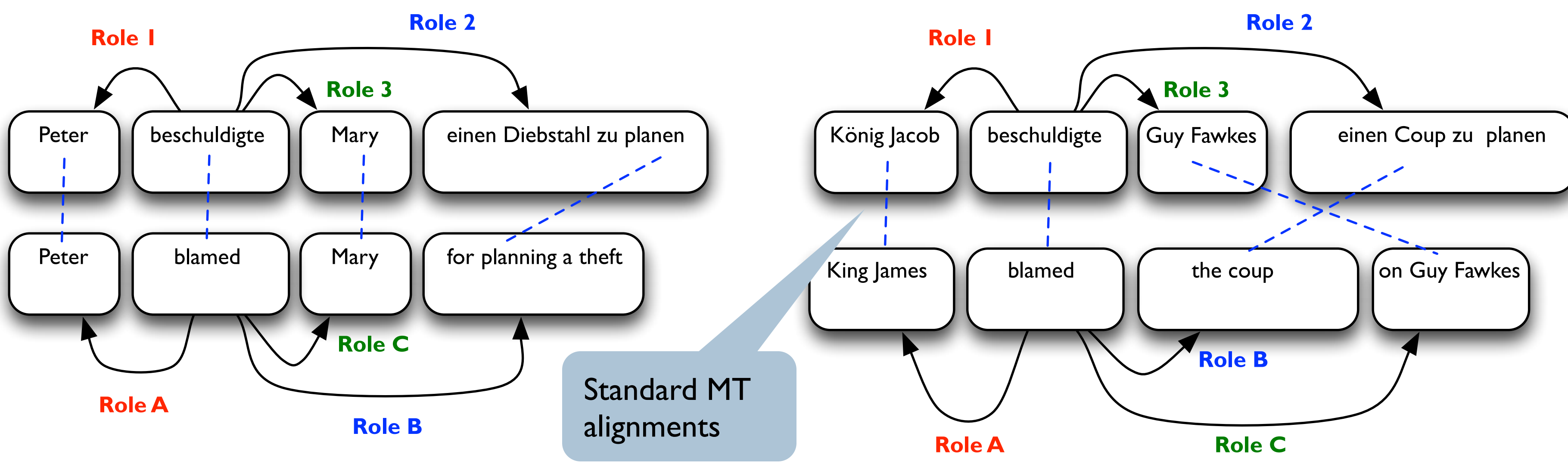


- However, the alternation does not transfer to German
  - Both forms are likely to have the same translation



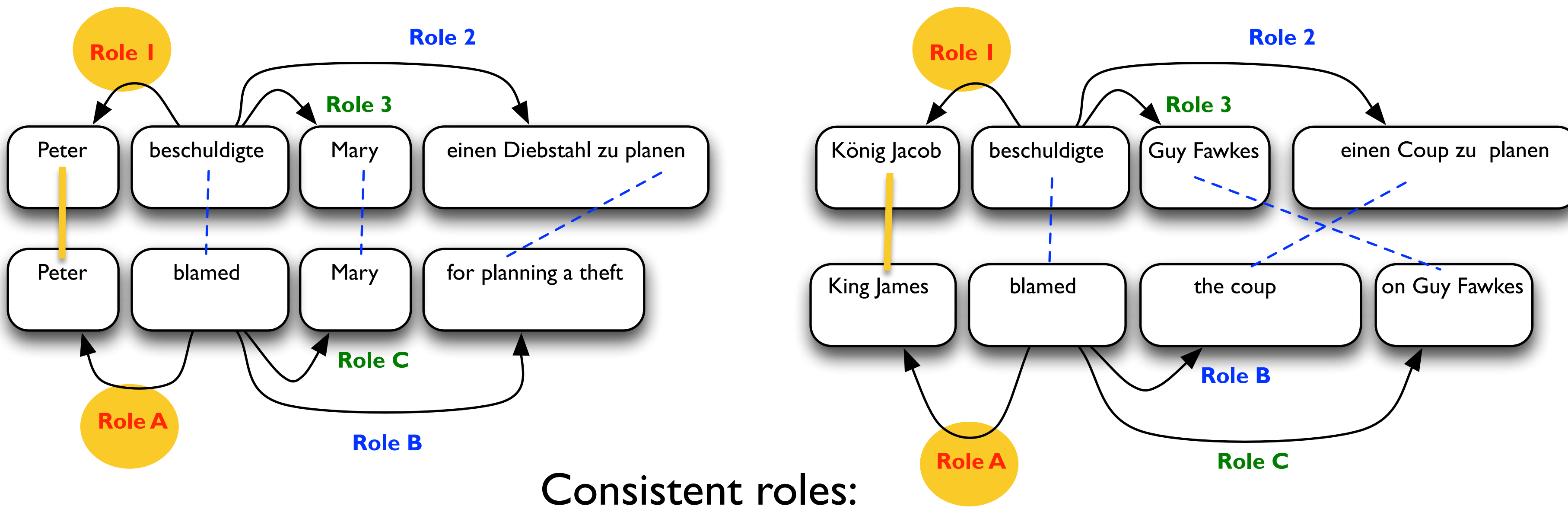
Crosslingual Induction of Semantic Roles

- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages



Crosslingual Induction of Semantic Roles

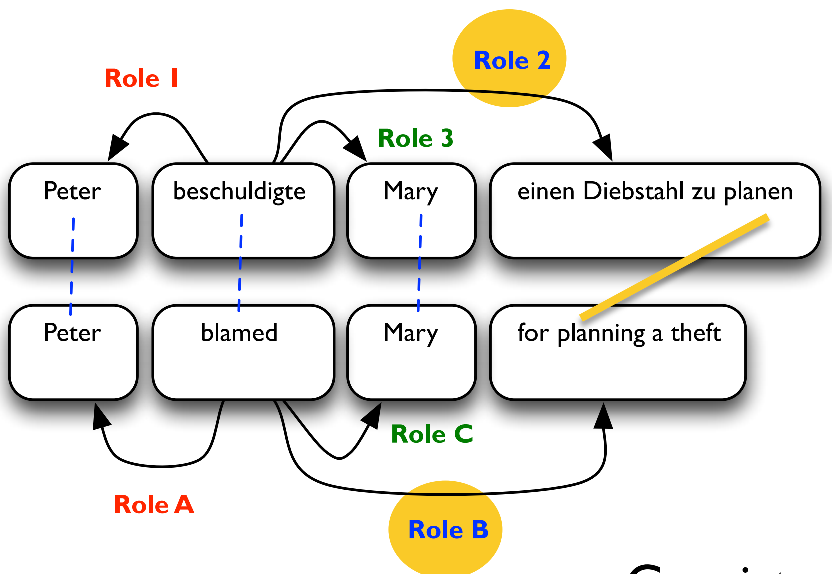
- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages

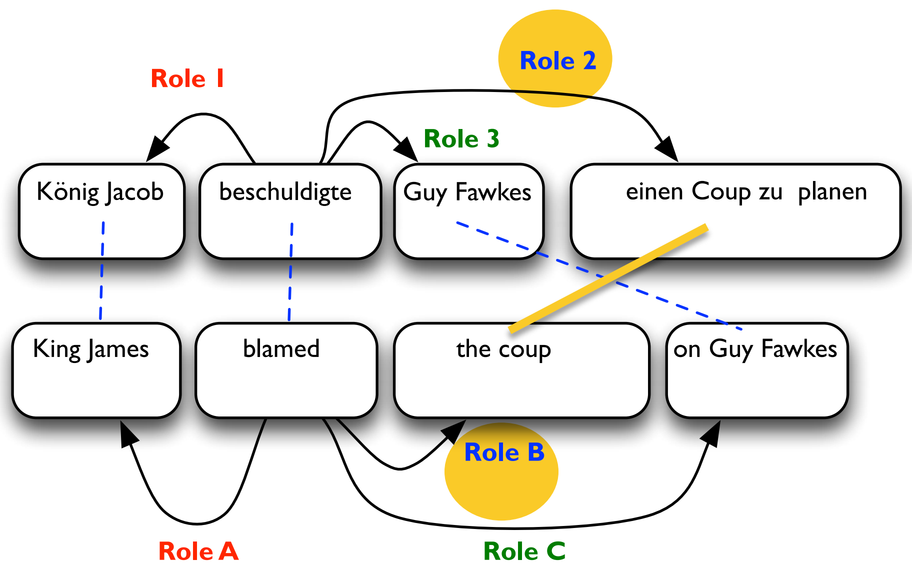


Consistent roles:
A to 1

Crosslingual Induction of Semantic Roles

- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages

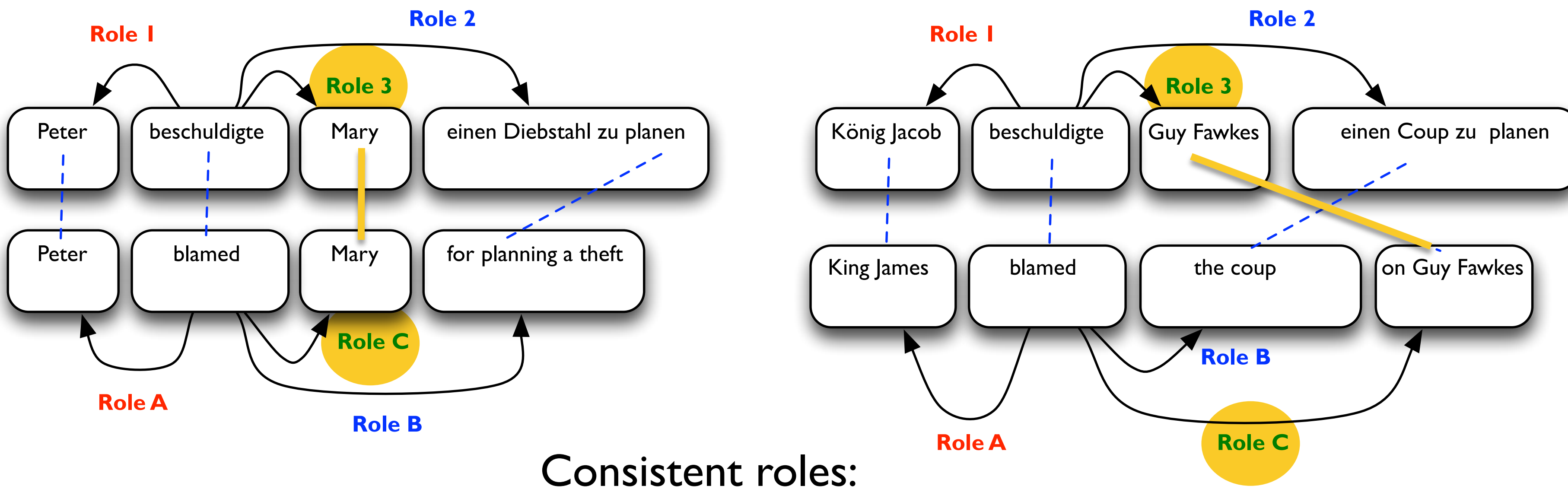




Consistent roles:
A to 1
B to 2

Crosslingual Induction of Semantic Roles

- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages

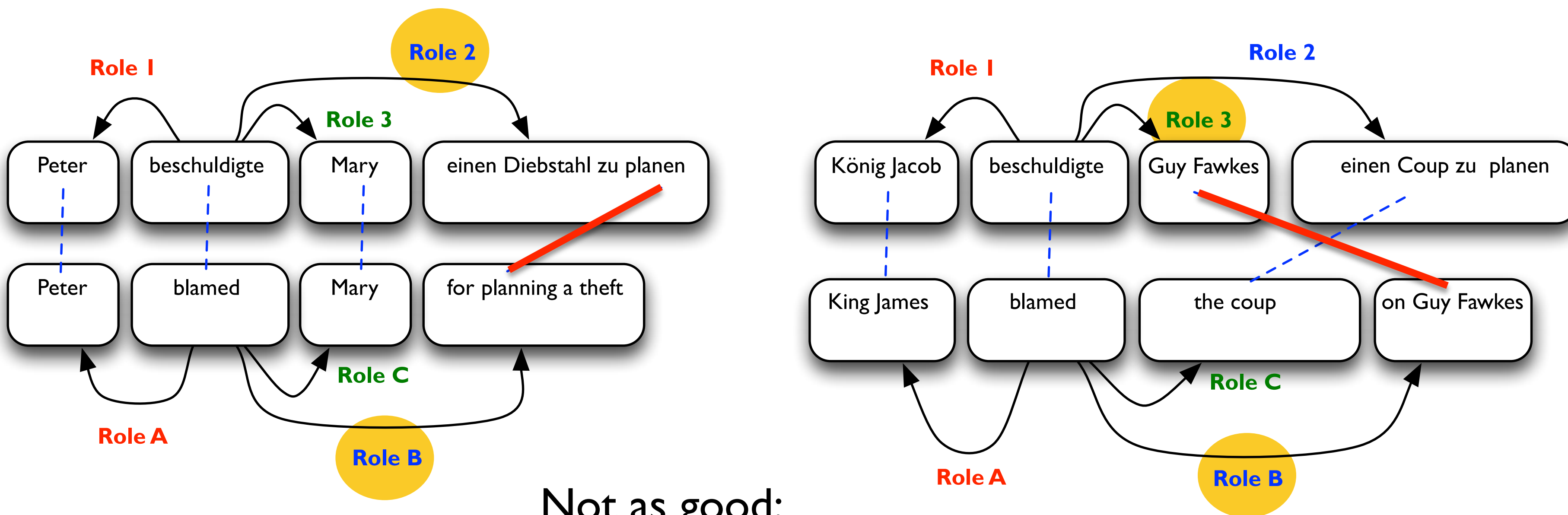


Consistent roles:
A to 1
B to 2
C to 3

Should be favored

Crosslingual Induction of Semantic Roles

- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages

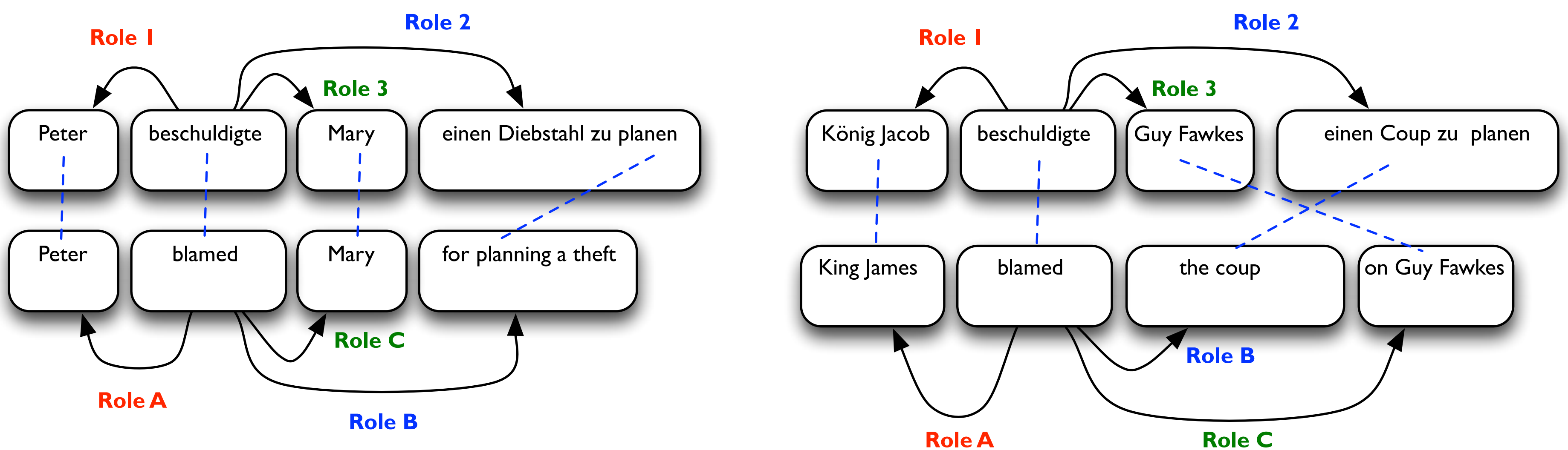


Not as good:
A to 1
B to 2 or 3
C to 3 or 2

Should be penalized

Crosslingual Induction of Semantic Roles

- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages

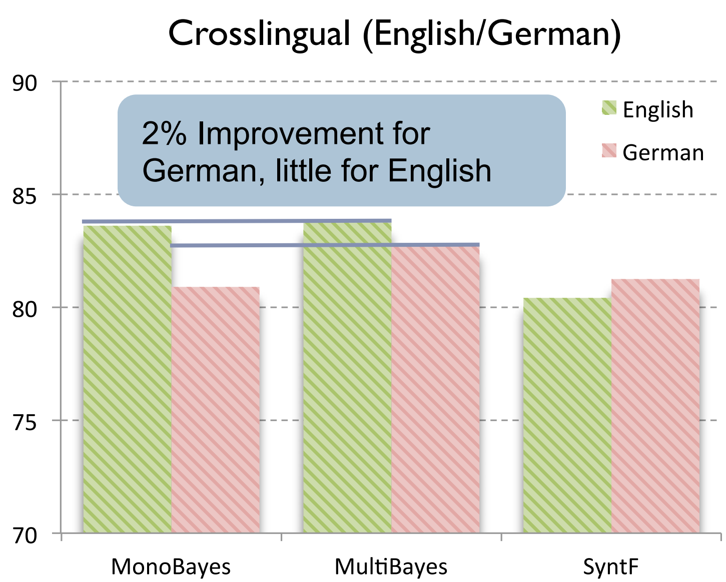


- In our example: roles induced for German will be transferred to English resulting in perfect accuracy on both languages
- Model extension (see Titov and Klementiev [ACL 2012]):
  - formulated as posterior regularization [Ganchev et al., 10, McCallum et al, 08].

Recall the Dipanjan's talk

Crosslingual Semantic Role Induction

- Experimental setup:
  - Induced jointly in two languages for predicates aligned in parallel data
  - Parallel data is used only to constrain the model to get fair comparison



Do we need unsupervised induction?

- Recall the Dipanjan's talk on Saturday:

> Crosslingual projection and (forms of) model transfer substantially outperform unsupervised induction of syntax / PoS tags

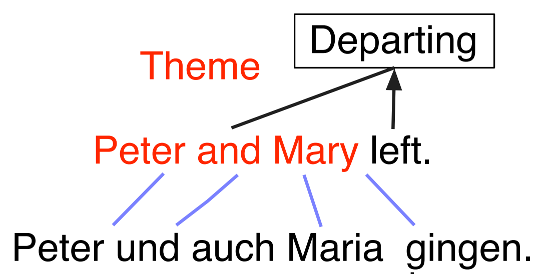
Do we need unsupervised induction?

- Recall the Dipanjan's talk on Saturday:

> Crosslingual projection and (forms of) model transfer substantially outperform unsupervised induction of syntax / PoS tags

- Annotation projection:
  - project annotation from the source language to the target language

[Pado and Lapata, 2005; Johansson and Nugues, 2006; Pado and Pitel, 2007; Tonelli and Pianta, 2008,…]



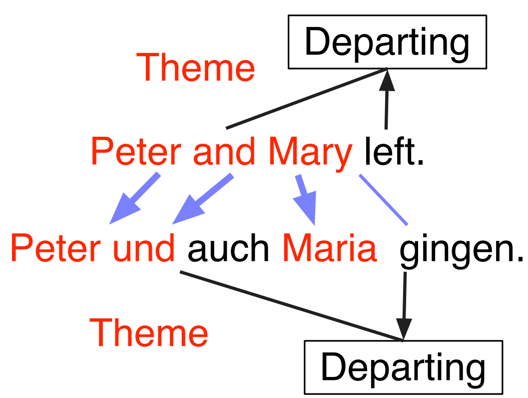
Do we need unsupervised induction?

- Recall the Dipanjan's talk on Saturday:

> Crosslingual projection and (forms of) model transfer substantially outperform unsupervised induction of syntax / PoS tags

- Annotation projection:
  - project annotation from the source language to the target language

[Pado and Lapata, 2005; Johansson and Nugues, 2006; Pado and Pitel, 2007; Tonelli and Pianta, 2008,…]



Do we need unsupervised induction?

- Recall the Dipanjan's talk on Saturday:

> Crosslingual projection and (forms of) model transfer substantially outperform unsupervised induction of syntax / PoS tags

- Annotation projection:
  - project annotation from the source language to the target language

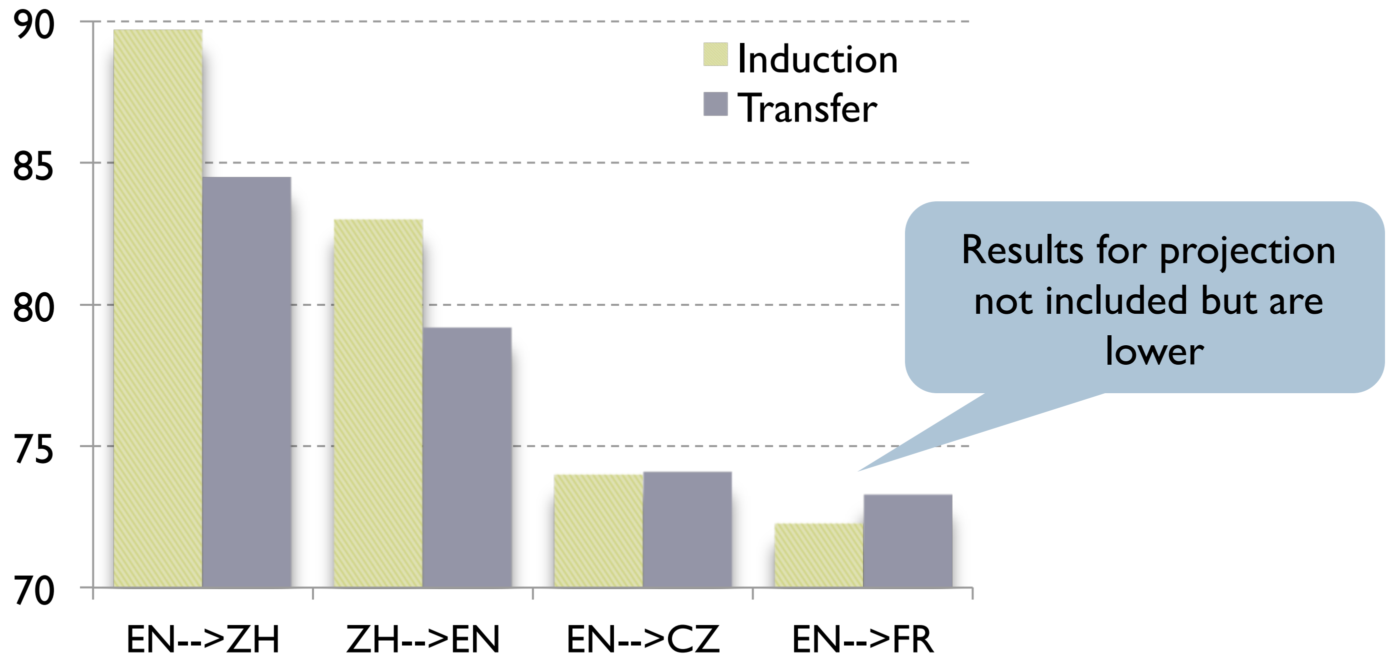
[Pado and Lapata, 2005; Johansson and Nugues, 2006; Pado and Pitel, 2007; Tonelli and Pianta, 2008,…]

- Model transfer:
  - apply a source SRL model to the target language (maybe with some adaptation)

[Kozhevnikov and Titov, 2013; Kozhevnikov and Titov, 2014]

[Kozhevnikov and Titov, 2013]

Induction vs. Transfer



The situation is quite different from the one for syntax / PoS tags

- Why?
  - divergences in semantic formalism across languages
  - semantics is more tied to lexical information so harder even for supervised methods

Outline

- Induction of events and their participants
  - unsupervised models of semantic roles
  - joint induction of frames and roles
  - cross-lingual extension and comparison with projection and transfer
- Induction of semantic representations of words (and phrases)
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, lexicon induction)

Why not clustering as before?

Clustering	Distributed (= Latent Features)
Cluster words into (hierarchical) clusters	Dense embedding
Words defined by cluster prototypes	

Clustering:

How to choose granularity?

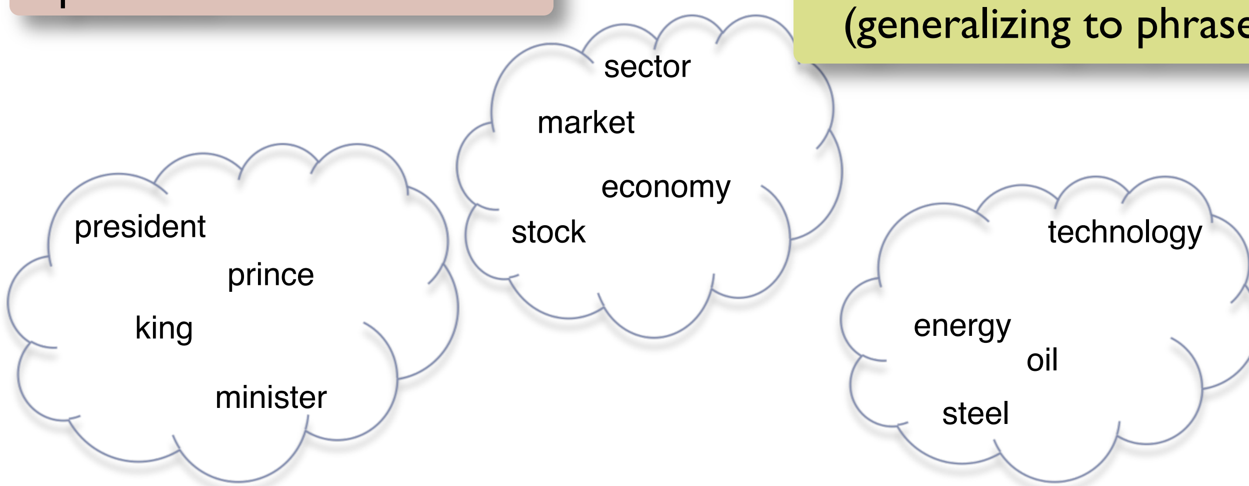
Many incompatible ways to cluster are often possible

Distributed (= Latent Features):

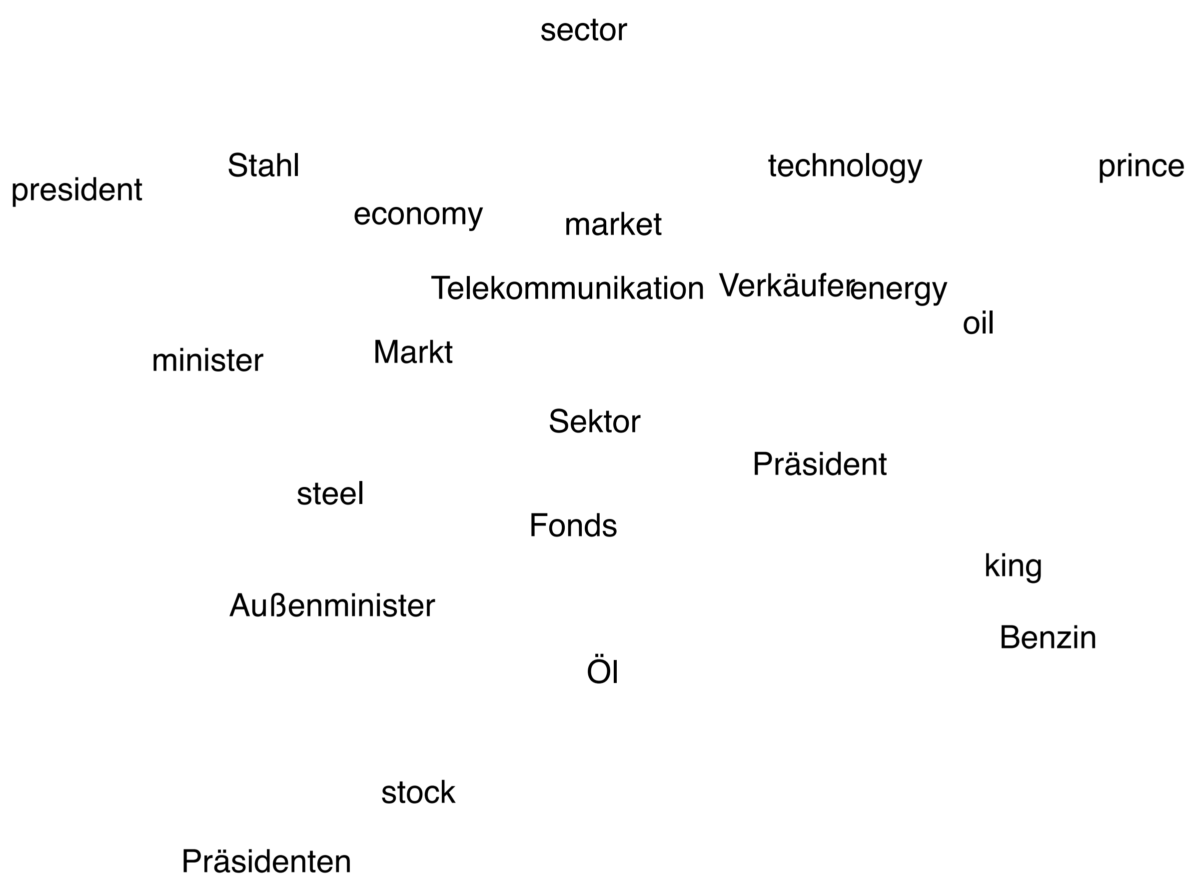
Can encode different levels of granularity

Can encode multiple incompatible clusterings (or multiple senses)

Easier to deal with compositionality (generalizing to phrases)

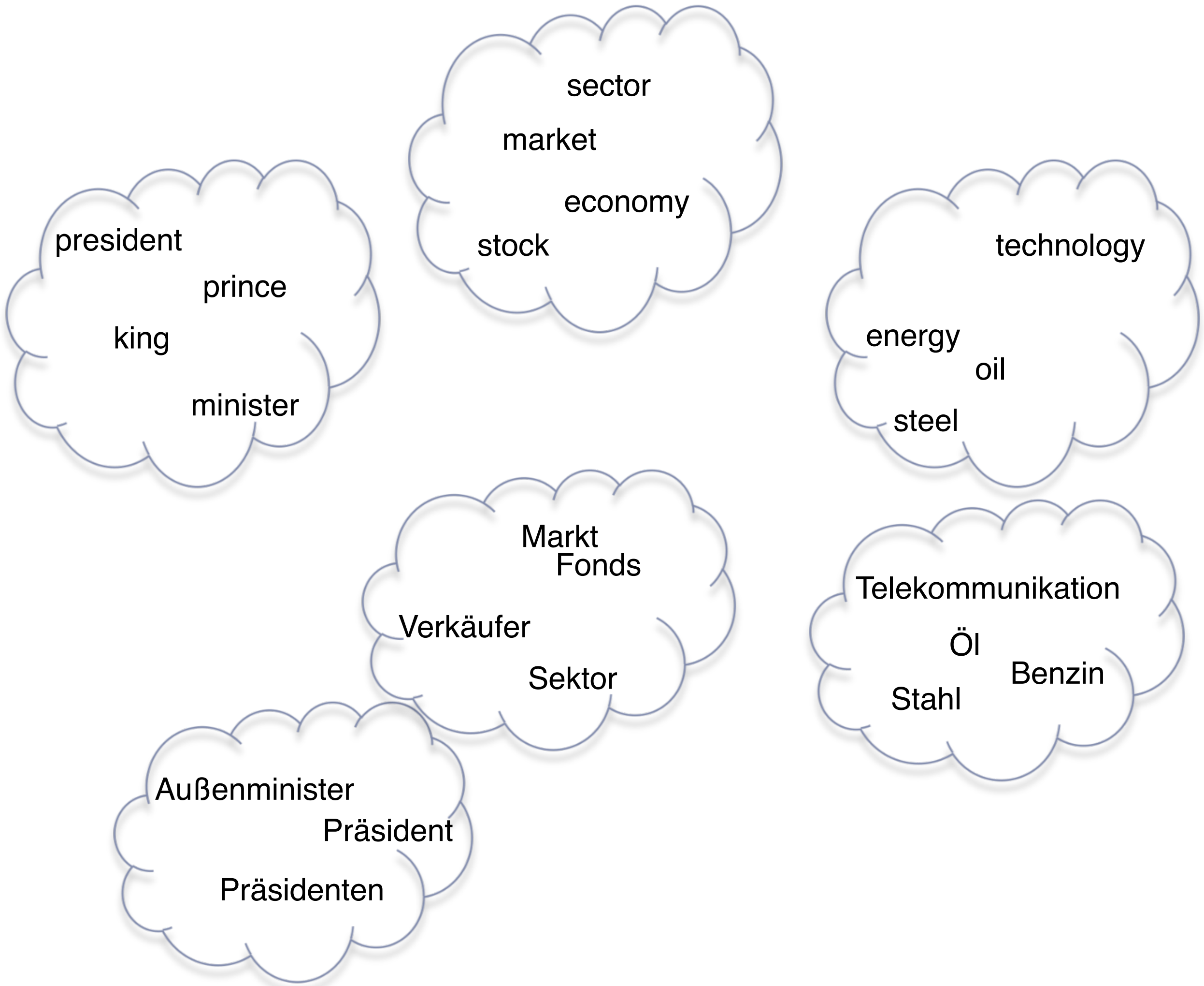


Summary of our Approach



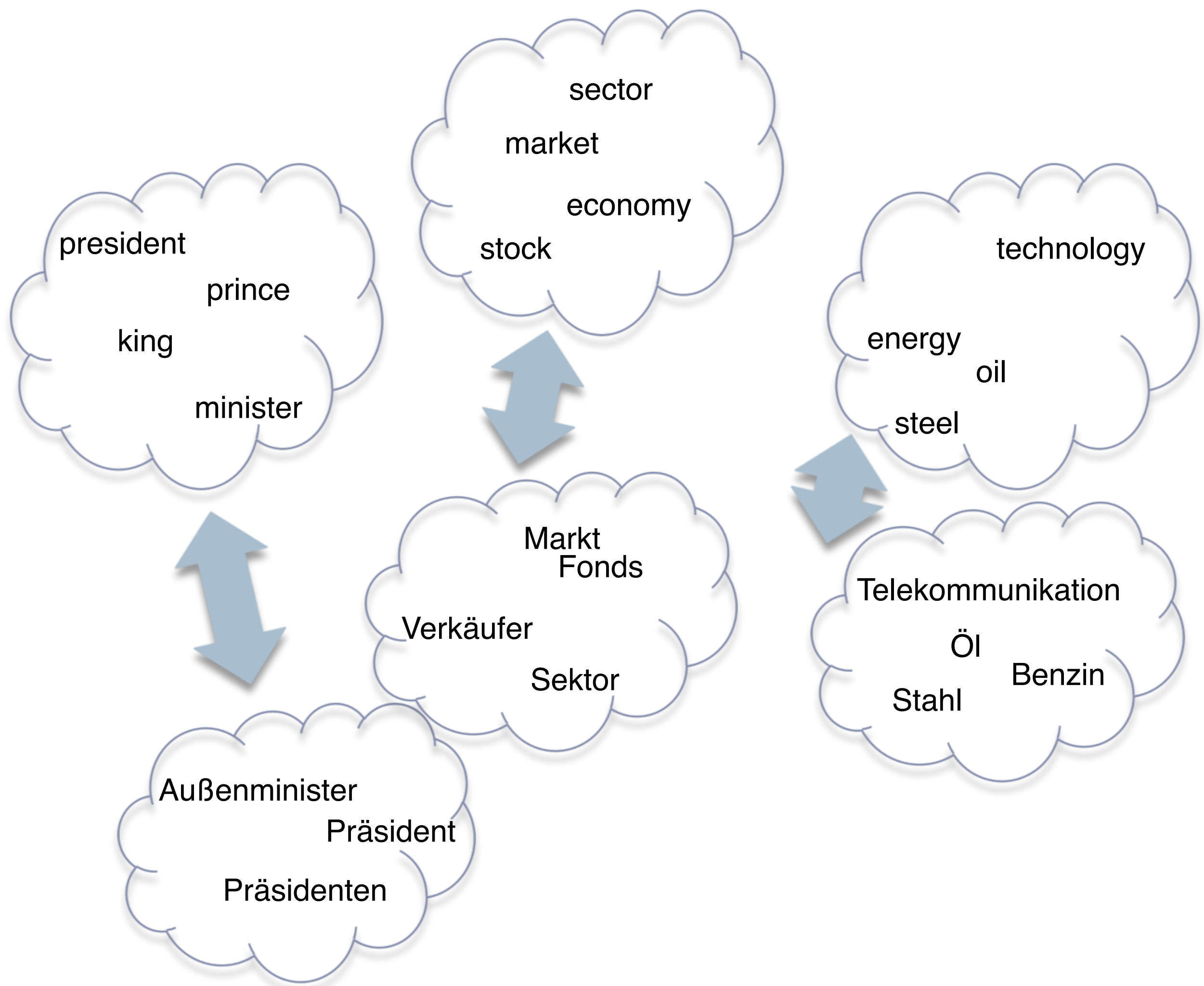
Summary of our Approach

- Use cheap monolingual data to induce the representation within each language



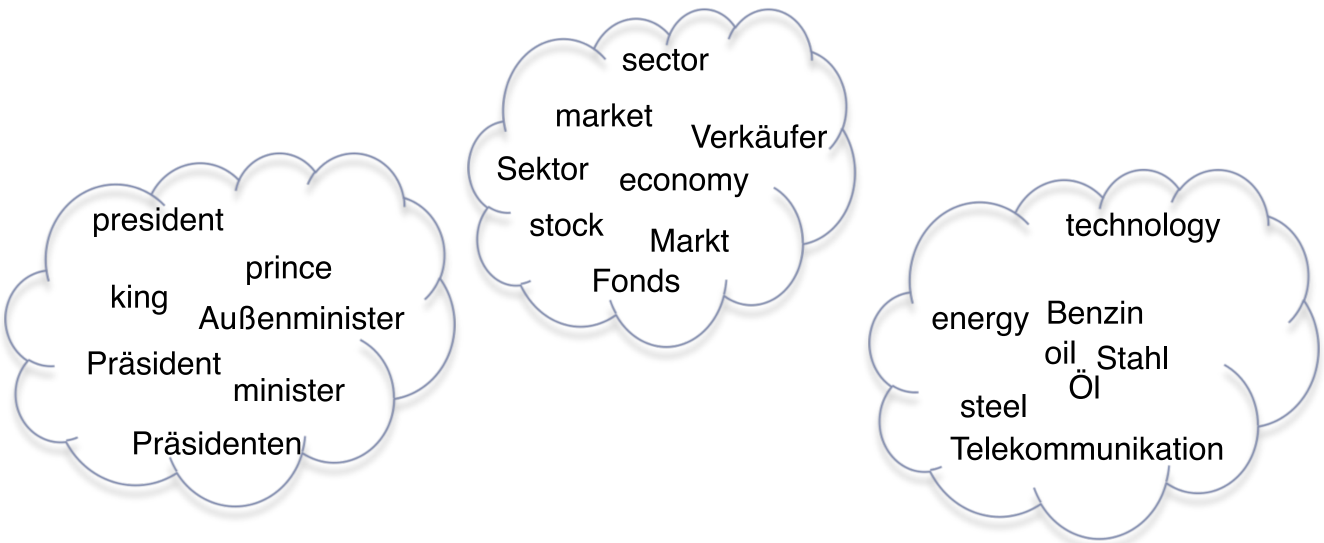
Summary of our Approach

- While using parallel data to bias representations to be similar for translated words



Summary of our Approach

- Semantically similar words are "close" to one another irrespective of language



- Treat it as multitask learning (MTL)
  - Treat words as individual tasks
  - Task relatedness is derived from co-occurrence statistics in bilingual parallel data

Background: Multitask Learning

- We consider a particular MTL setup [Cavallanti et al. (2010)]
- Consider K related tasks with a labeled dataset for each task k
- Learns a classifier (parameterized by $\boldsymbol{v_k}, k \in [1, K]$) for each task
- Minimizes the following objective:

Matrix A defines inter-task similarity

$$L(\boldsymbol{v}) = \sum_{k=1}^{K} L^{(k)}(\boldsymbol{v}_k) \;+\; \boxed{\frac{1}{2}\boldsymbol{v}^T(A \otimes I_m)\boldsymbol{v}}$$

Objectives for each individual task (e.g., likelihoods of each dataset)

Regularizer prefers "similar" parameters for related tasks

(Klementiev, Titov, Bhattarai, COLING 2012)

What do we take from MLT?

Idea: frame crosslingual distributed representation induction as multi-task learning

- We treat words in both languages as individual tasks
  - For each word, we learn a representations $\boldsymbol{c}_i \in \mathcal{R}^d$
- *A* will be defined by how often words align in parallel data
- We will take the multitask regularizer part of the objective

$$L(\boldsymbol{c}, \boldsymbol{\theta}) = \sum_{l=1}^{2} L^{(l)}(\boldsymbol{c}, \boldsymbol{\theta}_l) \quad + \quad \frac{1}{2}\boldsymbol{c}^T (A \otimes I_m)\boldsymbol{c}$$

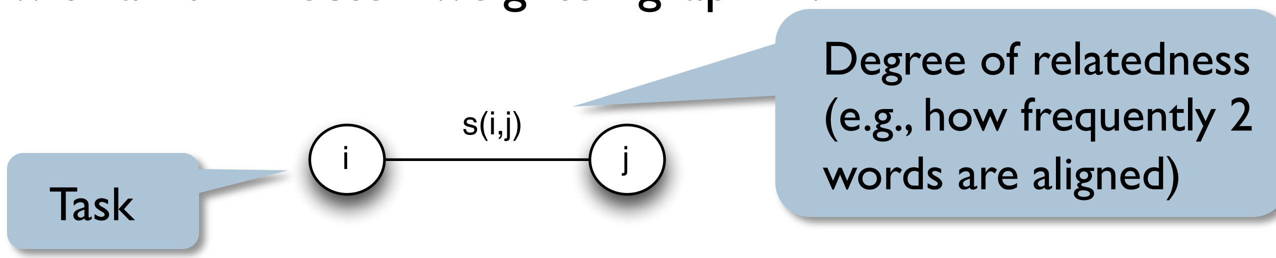
Loss function for a dataset in every language

Favors similar representations for frequently aligned words

- Applicable to any distributed representation induction set-up
  - We use neural probabilistic language model (Bengio et al, 2003)

How to encode relatedness?

- How can we encode prior knowledge of task (= word) relatedness into A?
- Represent tasks with an undirected weighted graph *H*:



- The graph *Laplacian L* is defined as:

$$L_{i,j}(H) = \begin{cases} \sum_{(i,k)\in E} s(i,k) & \text{if } i = j \\ -s(i,j) & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

- Interaction matrix is then defined as $A = I + L$
  - A^{-1} (crucial in learning) encodes the degree of relatedness between the tasks
  - *A* is invertible (*L* is positive semi-definite)

Qualitative Evaluation

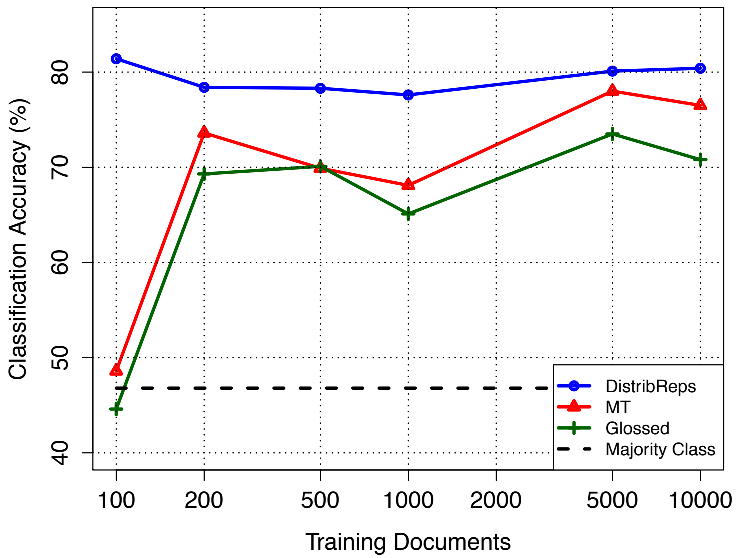
january		*president*		*said*	
en	**de**	**en**	**de**	**en**	**de**
january	januar	president	präsident	said	sagte
february	februar	king	präsidenten	reported	erklärte
november	november	hun	minister	stated	sagten
april	april	areas	staatspräsident	told	meldete
august	august	saddam	hun	declared	berichtete
march	märz	minister	vorsitzenden	stressed	sagt
june	juni	advisers	us-präsident	informed	ergänzte
december	dezember	prince	könig	announced	erklärten
july	juli	representative	berichteten	explained	teilt
september	september	institutional	außenminister	warned	berichteten

oil		*microsoft*		*market*	
en	**de**	**en**	**de**	**en**	**de**
oil	baumwolle	microsoft	microsoft	market	markt
car	kaffee	intel	intel	papers	marktes
energy	telekommunikation	instrument	chemikalien	side	fonds
air	tabak	chapman	endesa	economy	sektor
tobacco	rindfleisch	endesa	kabel	duration	laufzeit
steel	öl	distillates	hewlett-packard	sector	montreal
housing	benzin	pty	guinness	tobacco	verkäufer
cotton	stahl	hewlett-packard	dienste	montreal	papiere
insurance	strom	guinness	thomson	house	fracht
technology	milch	potash	exxon	pay	hersteller

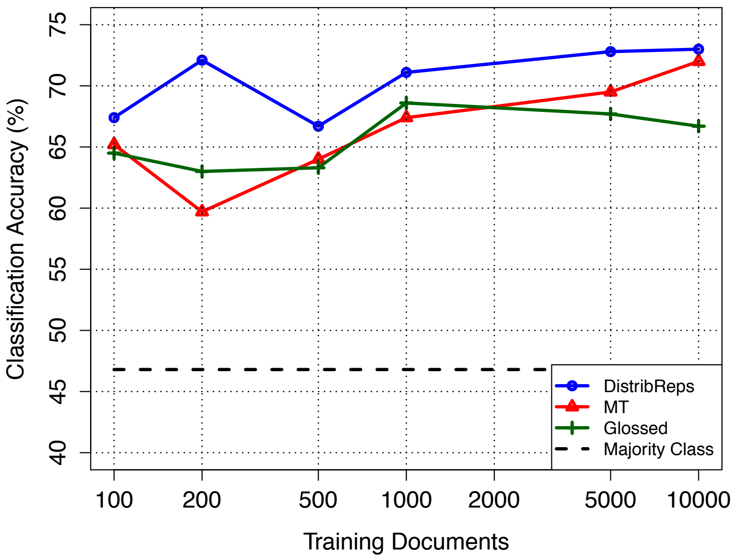
Crosslingual Document Classification

- Use distributed representations to train a classifier in one language (L1)
- Apply to the other language (L2) with *no* additional training (*DistribReps*)
- Baselines:
  - Train in L1, gloss test documents from L2 to L1 (*Glossed*)
  - Train in L1, translate (phrase-based MT) test documents in L2 to L1 (*MT*)

No training data in L2!!!



Train: en, Test: de



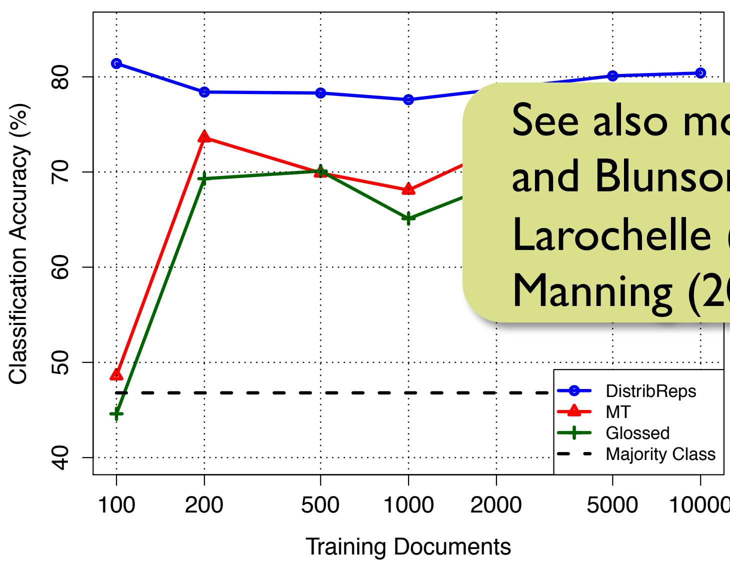
Train: de, Test: en

(Klementiev, Titov, Bhattarai, COLING 2012)

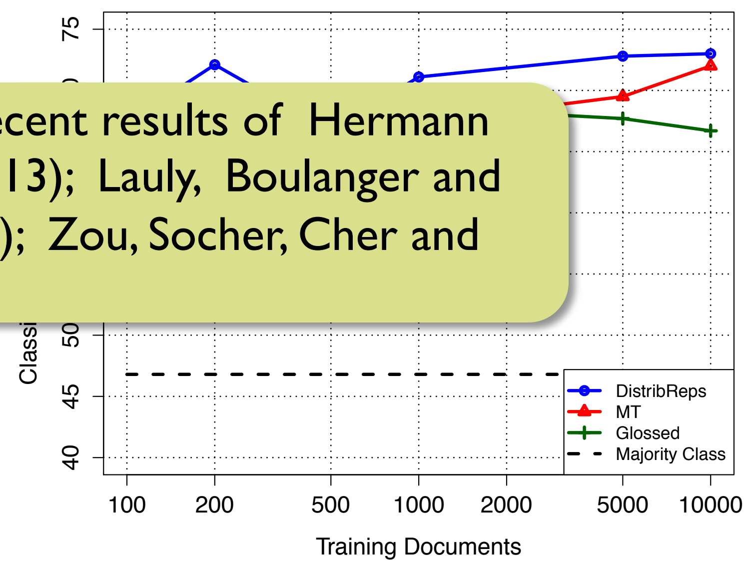
Crosslingual Document Classification

- Use distributed representations to train a classifier in one language (L1)
- Apply to the other language (L2) with *no* additional training (*DistribReps*)
- Baselines:
  - Train in L1, gloss test documents from L2 to L1 (*Glossed*)
  - Train in L1, translate (phrase-based MT) test documents in L2 to L1 (*MT*)

No training data in L2!!!



Train: en, Test: de



Train: de, Test: en

See also more recent results of Hermann and Blunsom (2013); Lauly, Boulanger and Larochelle (2014); Zou, Socher, Cher and Manning (2013)

Conclusions

- We believe that unsupervised induction and its semi-supervised extensions are a very promising direction

- Crosslingual learning
  - Enforcing agreement using parallel data

- Ongoing work: beyond frames semantics:
  - Learning how events are organized in more complex activities (Frermann et al., 2014; Modi and Titov, 2014)

- Many questions remaining
  - more expressive models of alternations;
  - going beyond sentences;

  …

This work is partially supported by a Google Research Award. Also thanks to Manfred Pinkal, Alexis Palmer, Ryan McDonald, Caroline Sporleder

(Some) References

B. O'Connor. 2013. Learning Frames from Text with an Unsupervised Latent Variable Model. CMU TR.

D. Das, N. Schneider, D. Chen and N. Smith. 2010. Probabilistic Frame-Semantic Parsing. NAACL.

L. Frermann, I. Titov and M. Pinkal. 2014. A Hierarchical Bayesian Model for Unsupervised Induction of Script Knowledge. EACL.

Grenager and C. Manning. 2006. Unsupervised Discovery of a Statistical Verb Lexicon. EMNLP.

D. Kawahara, D. Peterson, O. Popescu, M. Palmer. 2014. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. EACL.

A. Klementiev, I. Titov and B. Bhattarai. 2012. Inducing Crosslingual Distributed Representations of Words. COLING.

M. Kozhevnikov and I. Titov. 2013. Crosslingual Transfer of Semantic Role Models. ACL.

M. Kozhevnikov and I. Titov. 2014. Crosslingual Model Transfer Using Feature Representation Projection. ACL Short.

J. Lang and M. Lapata. 2010. Unsupervised induction of semantic roles. ACL.

J. Lang and M. Lapata. 2011b. Unsupervised semantic role induction via split-merge clustering. ACL,.

J. Lang and M. Lapata. 2011a. Unsupervised semantic role induction with graph partitioning. EMNLP.

A. Modi, I. Titov and A. Klementiev, 2012. Unsupervised Induction of Frame-Semantic Representations. ILS Workshop, NAACL.

S. Pado and M. Lapata. 2005. Cross-linguistic Projection of Role-Semantic Information. EMNLP.

S. Pado and M. Lapata. 2008. Crosslingual annotation projection for semantic roles. Journal of Artificial Intelligence Research.

S. Pado and M. Lapata. 2006. Optimal Constituent Alignment with Edge Covers for Semantic Projection. ACL.

S. Pado and G. Pitel. 2007. Annotation précise du français en sémantique de rôles par projection cross-linguistique. TALN. A. Palmer and C. Sporleder. 2010. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. COLING.

L. van deer Plas, P. Merlo, and J. Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. ACL.

H. Poon and P. Domingos. 2008. Unsupervised Semantic Parsing. EMNLP.

I. Titov and A. Klementiev. 2011. A Bayesian Model for Unsupervised Semantic Parsing. ACL.

I. Titov and A. Klementiev. 2012a. A Bayesian Approach to Unsupervised Semantic Role Induction. EACL.

I. Titov and A. Klementiev. 2012b. Crosslingual Induction of Semantic Roles, ACL.

I. Titov and A. Klementiev. 2012c. Semi-supervised Semantic Role Labeling: Approaching from an Unsupervised Perspective. COLING.