Inducing Cross-Lingual Semantic Representations of Words, Phrases and Events

Ivan Titov

Joint work with Alex Klementiev, Binod Bhattarai, Chris Callison-Burch and Ann Irvine

CLUSTER OF EXCELLENCE

UNIVERSITÄT DES SAARLANDES

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] 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 induction of the frame-semantic representations
  - a Bayesian model and approximate inference
  - cross-lingual extension
  - evaluation (QA, PropBank)
- Induction of semantic representations of words and phrases
  - distributed representations vs. clustering
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, translation scoring)

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, …

Syntactic-Semantic 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:

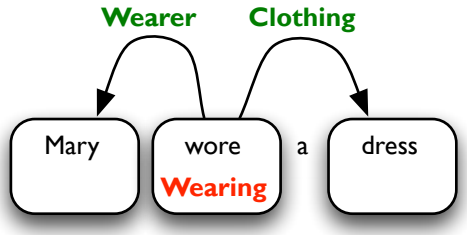
AGENT – an initiator/doer in the event [Who?]

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

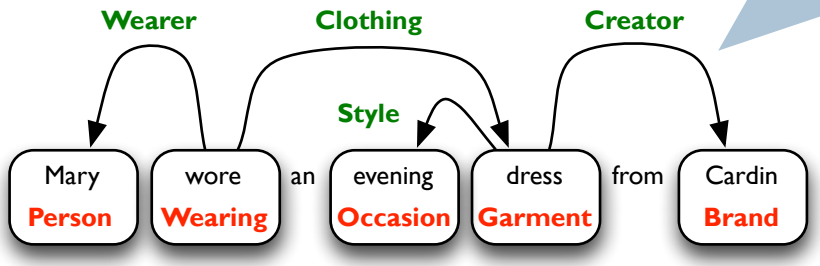
Our task

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

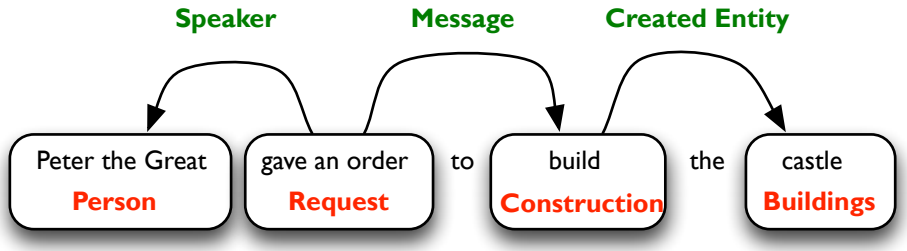


- Arguments often evoke their own frames

For simplicity we assume that all of them evoke frames



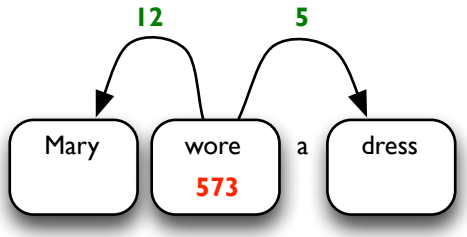
- Arguments and predicates often expressed by multiword expressions



Induce these representations automatically from unannotated texts

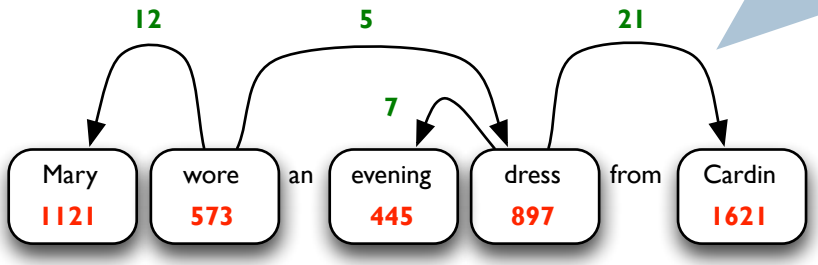
Our task

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

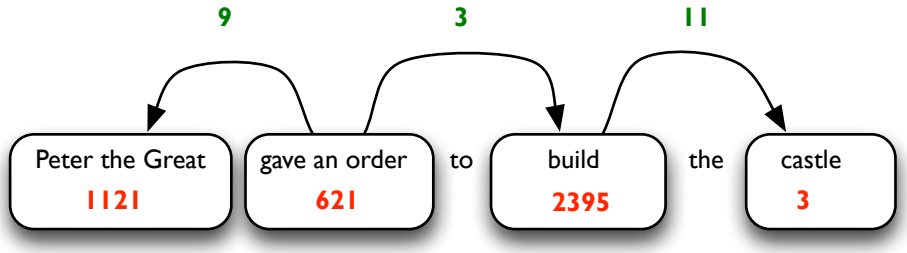


- Arguments often evoke their own frames

For simplicity we assume that all of them evoke frames



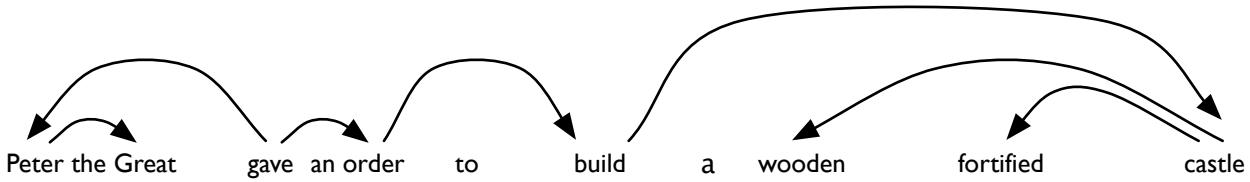
- Arguments and predicates often expressed by multiword expressions



Induce these representations automatically from unannotated texts

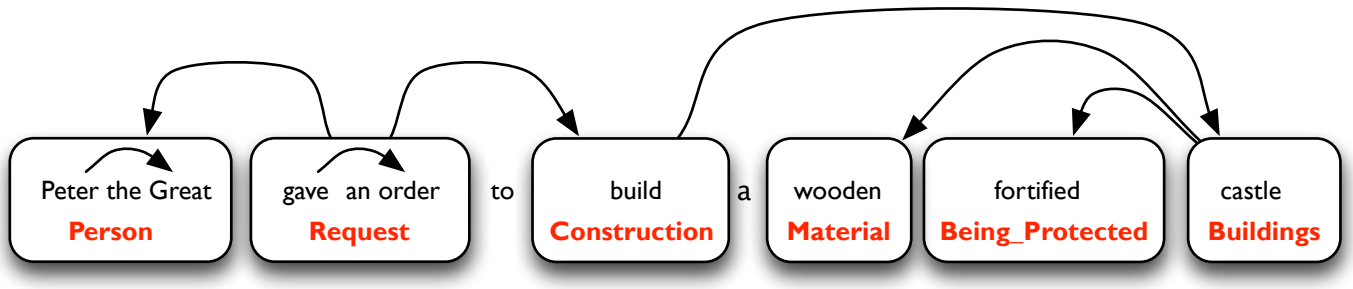
Induction of Frame-Semantic Information

- 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)

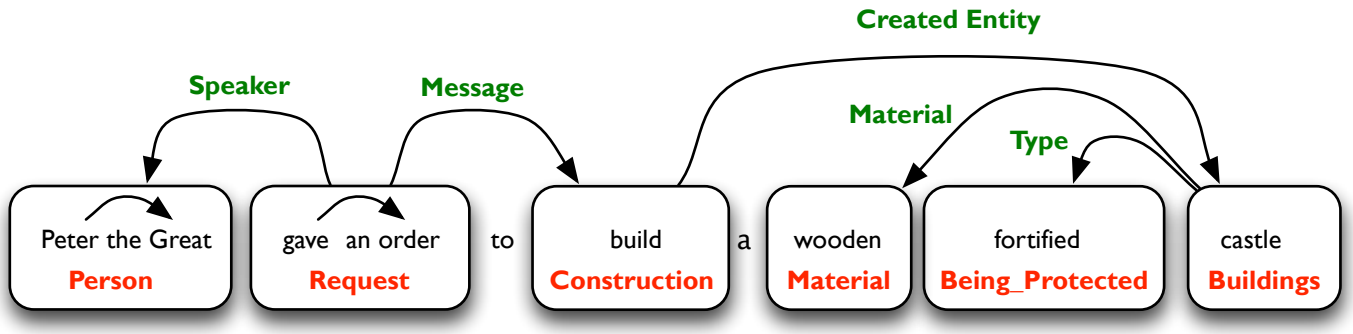


Induction of Frame-Semantic Inform

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

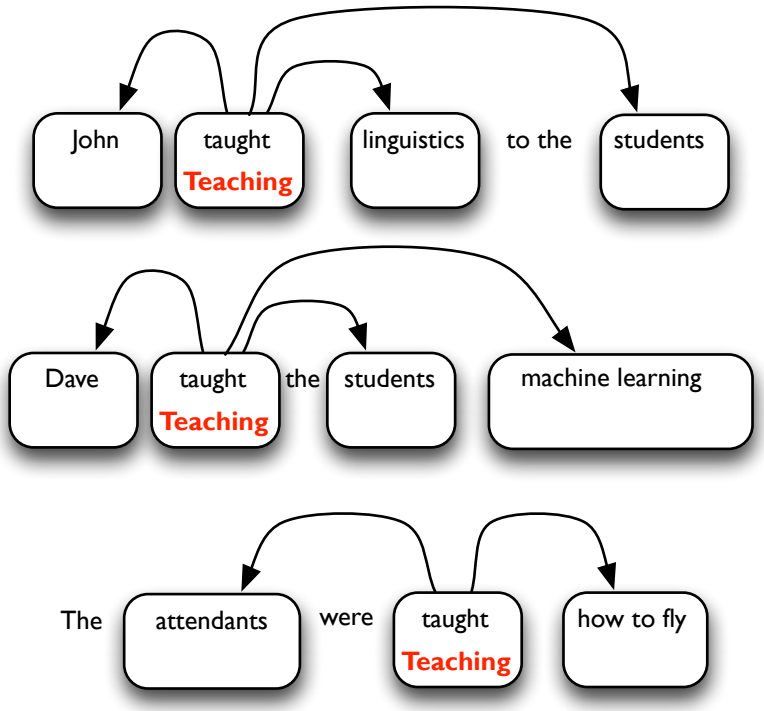
Induction of Semantic Classes: Definition

- 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

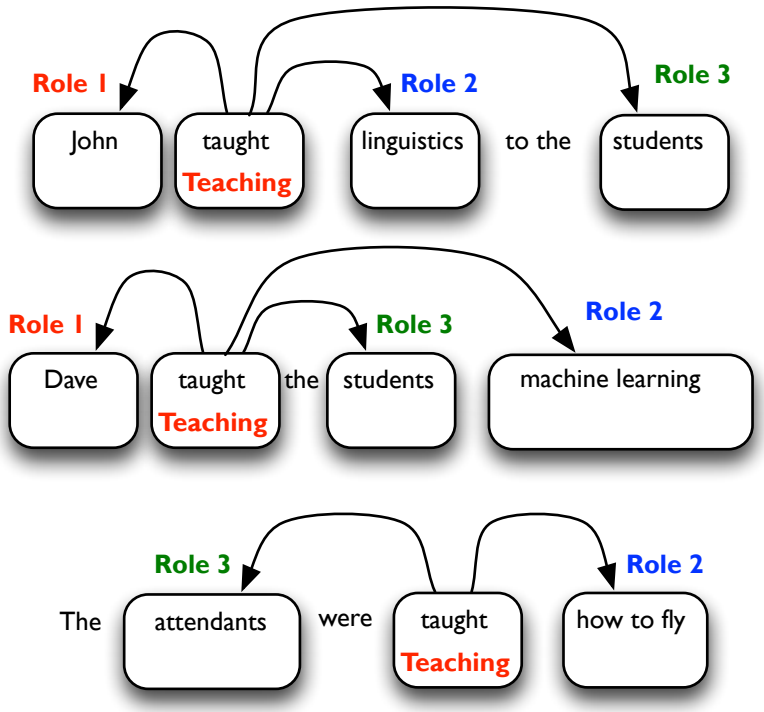
Induction of Semantic Roles: Definition

- 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>



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>



We need to "color" them

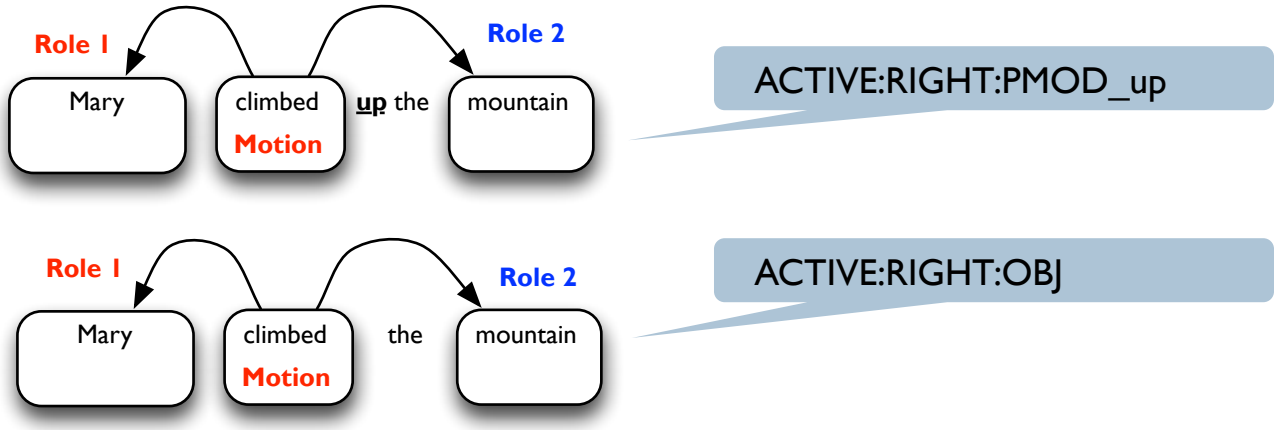
- The search space is huge – in realistic datasets frequents semantic classes appear tens of thousands times

Argument Keys

- We identify arg occurrences with syntactic signatures (argument keys)

  (as in Lang and Lapata [2011])

- E.g., some simple alternations like locative preposition drop



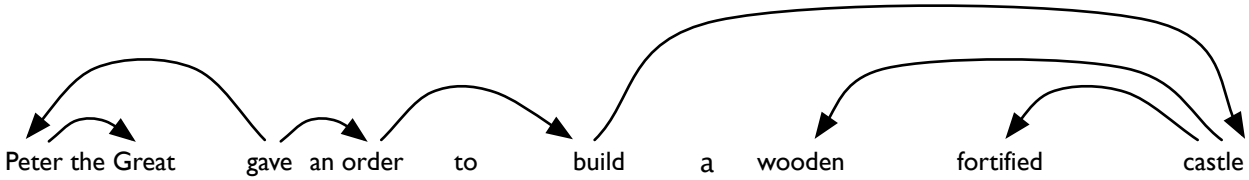
- Argument keys are designed so that to map mostly to a single role
- Instead of clustering occurrences we cluster argument keys
- Here, we would cluster ACTIVE:RIGHT:OBJ and ACTIVE:RIGHT:PMOD_up together
  - More complex alternations require multiples pairs of arg keys clustered

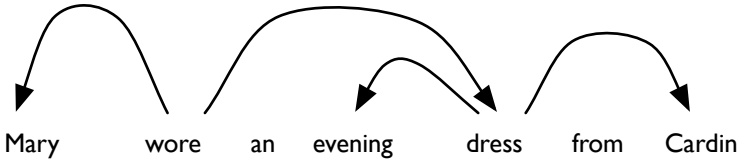
Outline

- Induction of events and their participants
  - unsupervised induction of the frame-semantic representations
  - a Bayesian model and approximate inference
  - cross-lingual extension
  - evaluation (QA, PropBank)
- Induction of semantic representations of words and phrases
  - distributed representations vs. clustering
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, translation scoring)

Inducing Semantics

- Given a (large) collection of sentences annotated with (transformed) syntactic dependencies $\{x_i\}_{i=1}^{n}$

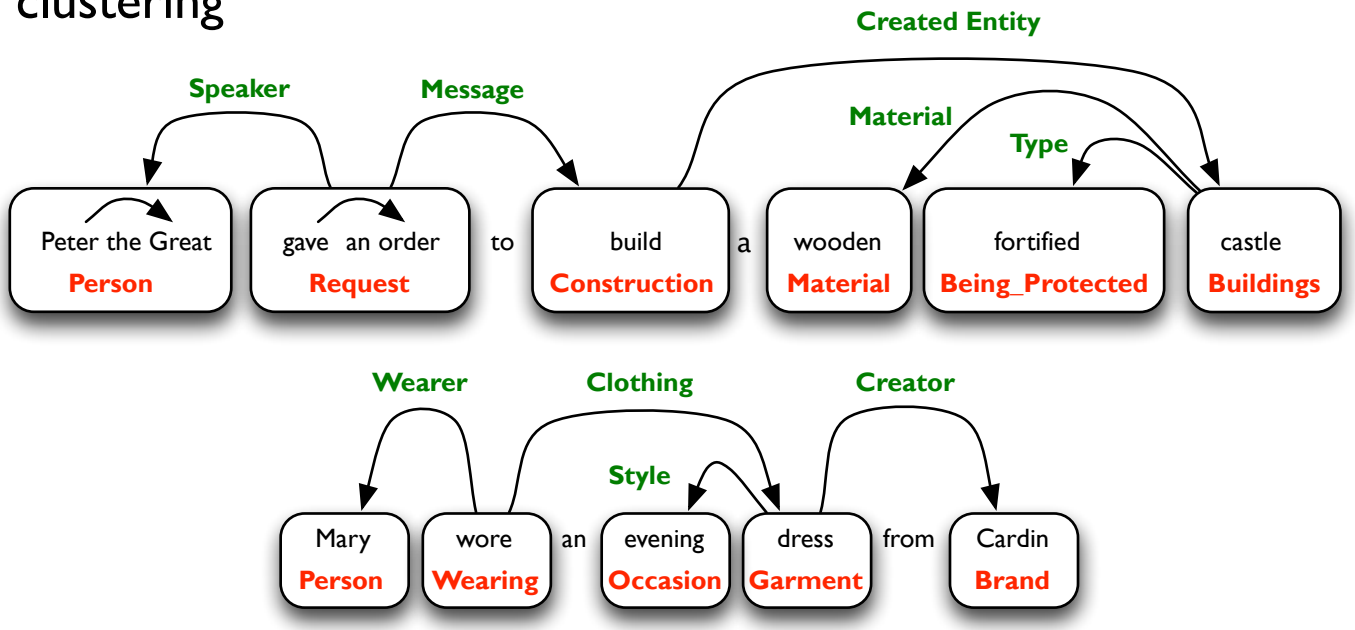




...

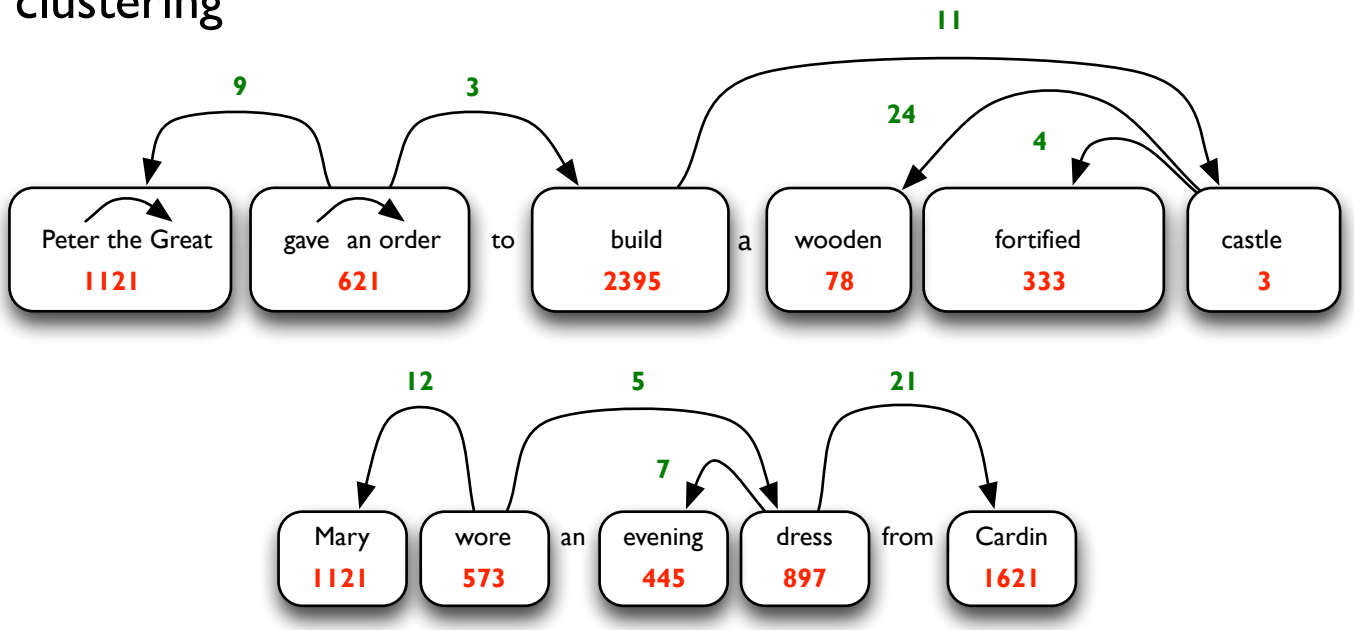
Inducing Semantics

- Given a (large) collection of sentences annotated with (transformed) syntactic dependencies $\{x_i\}_{i=1}^n$
- We want to induce its semantics $\{m_i\}_{i=1}^n$, i.e. its segmentation and clustering



Inducing Semantics

- Given a (large) collection of sentences annotated with (transformed) syntactic dependencies $\{x_i\}_{i=1}^n$
- We want to induce its semantics $\{m_i\}_{i=1}^n$, i.e. its segmentation and clustering



We use NP Bayes to induce the number of components (clusters, fragment sizes) required to explain the data

(Titov and Klementiev, ACL 2011, EACL 2012)

Modeling assumptions

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

(2) Each semantic class can be verbalized as a sparse distribution over lexemes or syntactic tree fragments

(3) Semantically-similar predicates have the same linking between syntax and semantics

(4) The same semantic role rarely appears twice

(5) Argument key clusterings for different predicates are related

The (Simplified) Model

Draw semantic class for root

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

The (Simplified) Model

Draw semantic class for root

$for\ each\ sentence:$

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

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

Request

GenSemClass(c)

$s \sim \phi_c$

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

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

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

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

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

The (Simplified) Model

Draw semantic class for root

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

Draw synt/lex realization

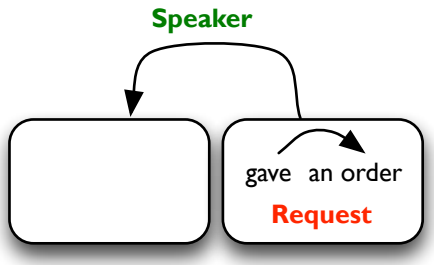
$$
\begin{aligned}
&\mathbf{GenSemClass}(c) \\
&\quad s \sim \phi_c \\
&\quad for\ each\ role\ t = 1, \ldots, T: \\
&\quad if\ [n \sim \psi_{c,t}] = 1: \\
&\quad\quad \mathbf{GenArgument}(c, t) \\
&\quad\quad while\ [n \sim \psi^{+}_{c,t}] = 1: \\
&\quad\quad\quad \mathbf{GenArgument}(c, t)
\end{aligned}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

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

The (Simplified) Model

Draw semantic class for root

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



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

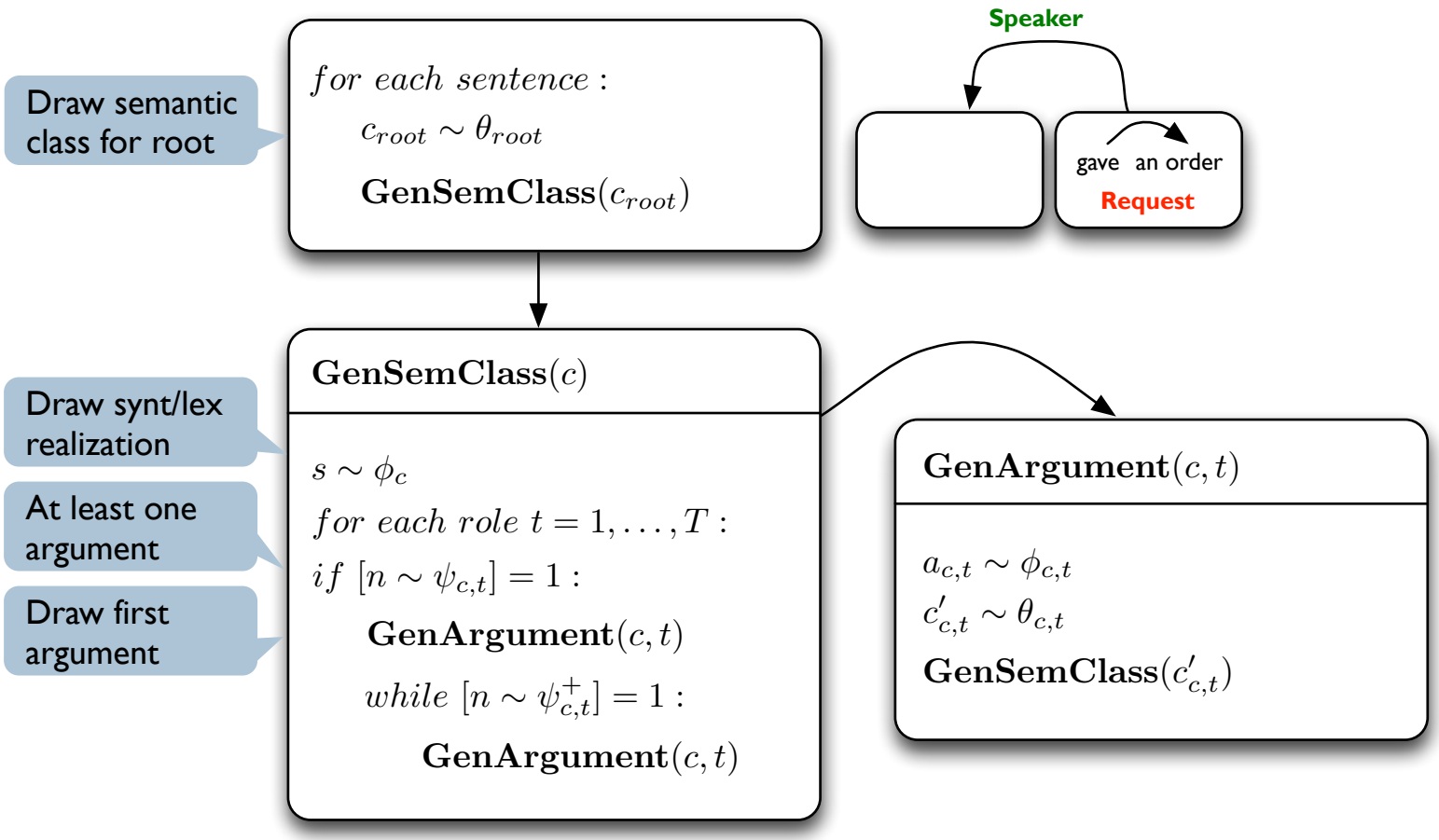
Draw synt/lex realization

$$s \sim \phi_c$$

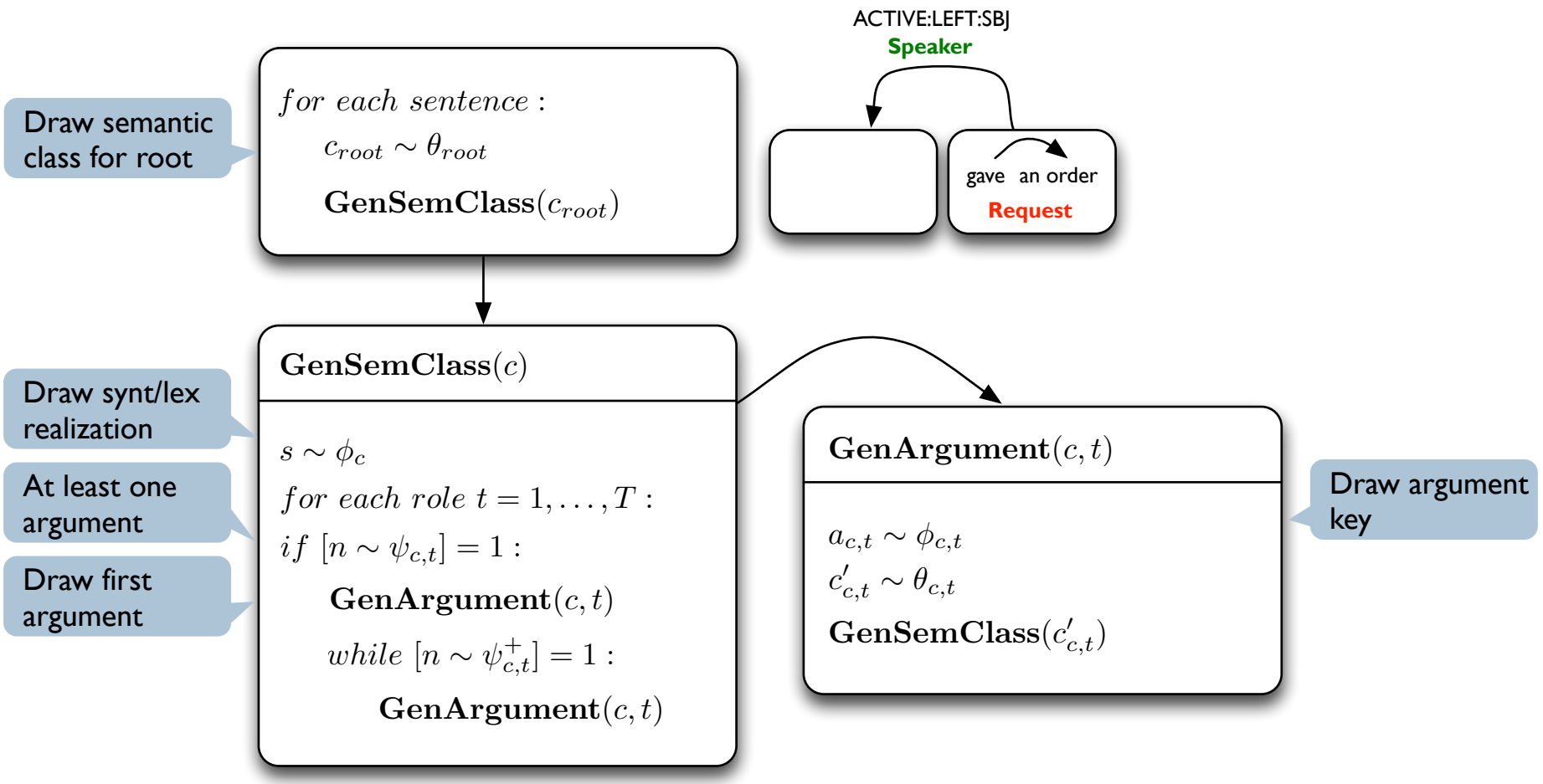
At least one argument

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

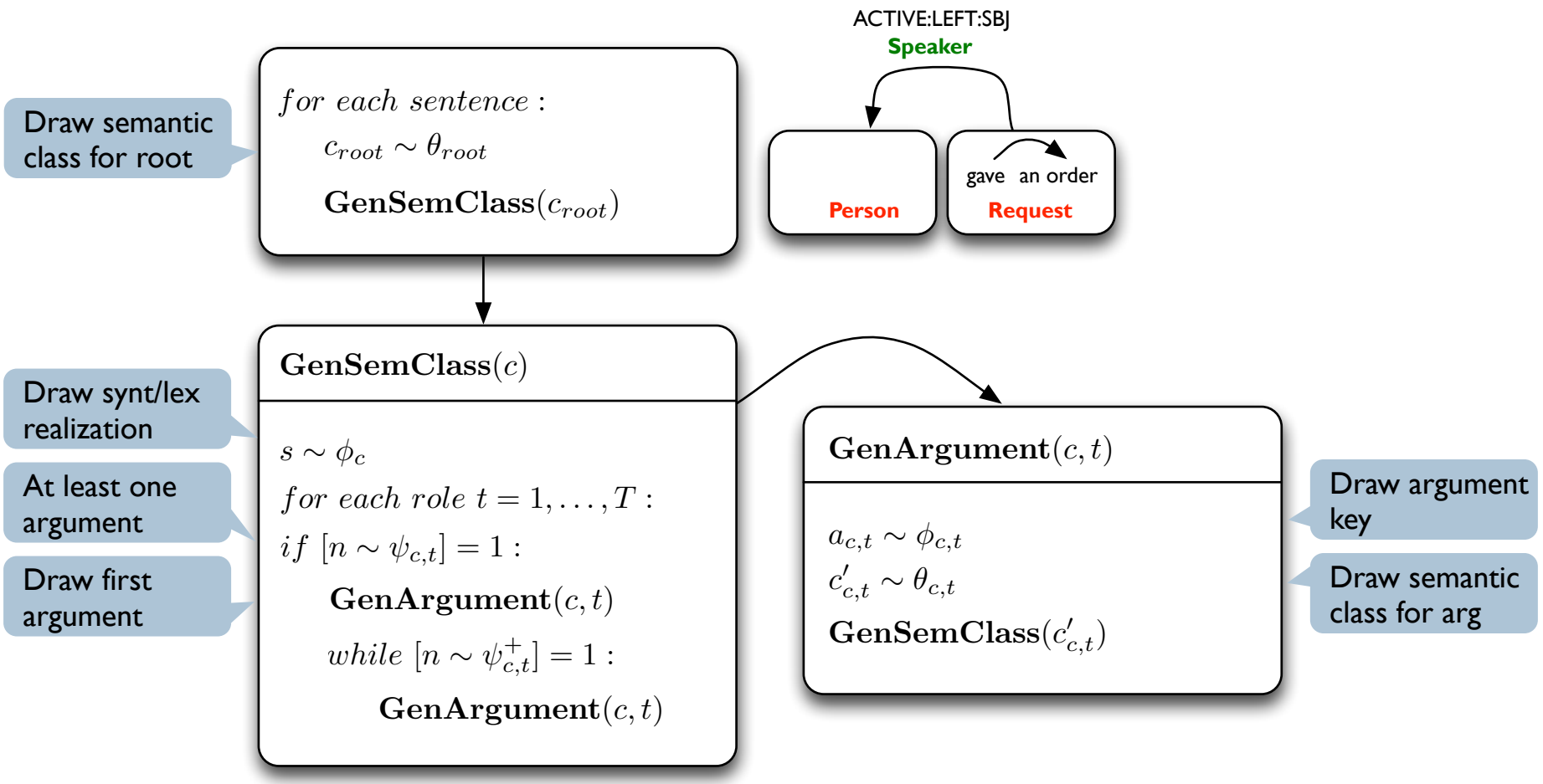
The (Simplified) Model



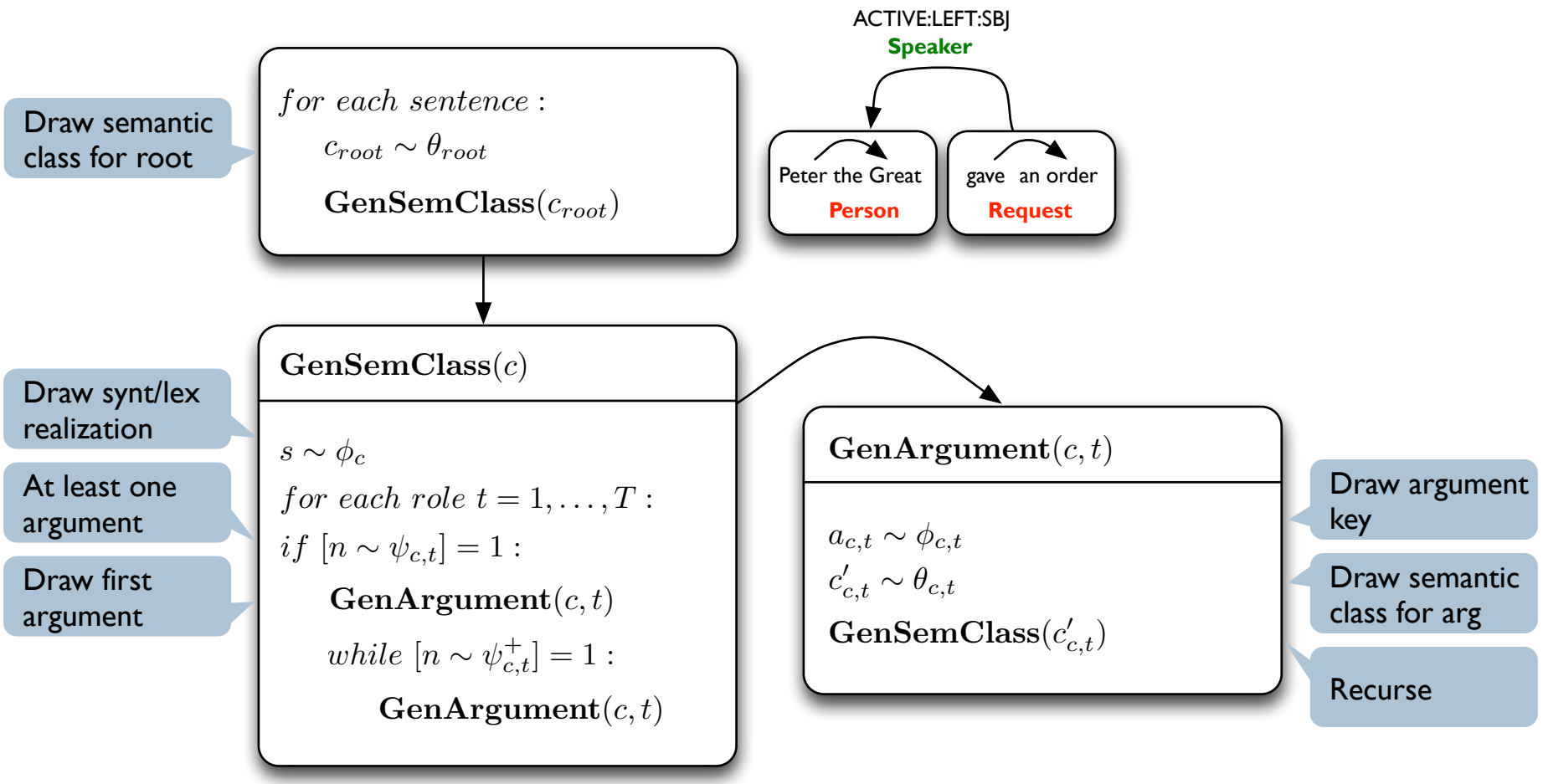
The (Simplified) Model



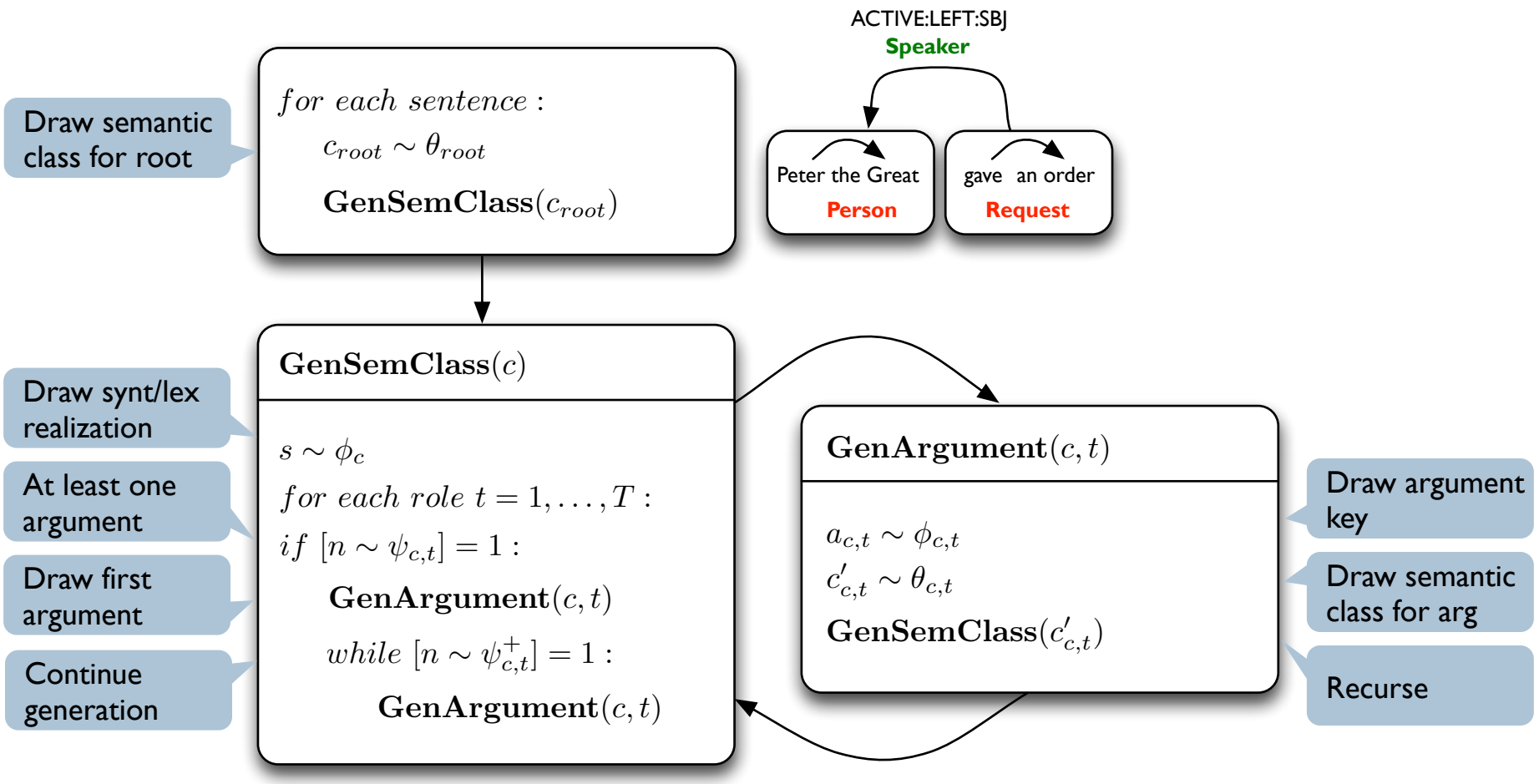
The (Simplified) Model



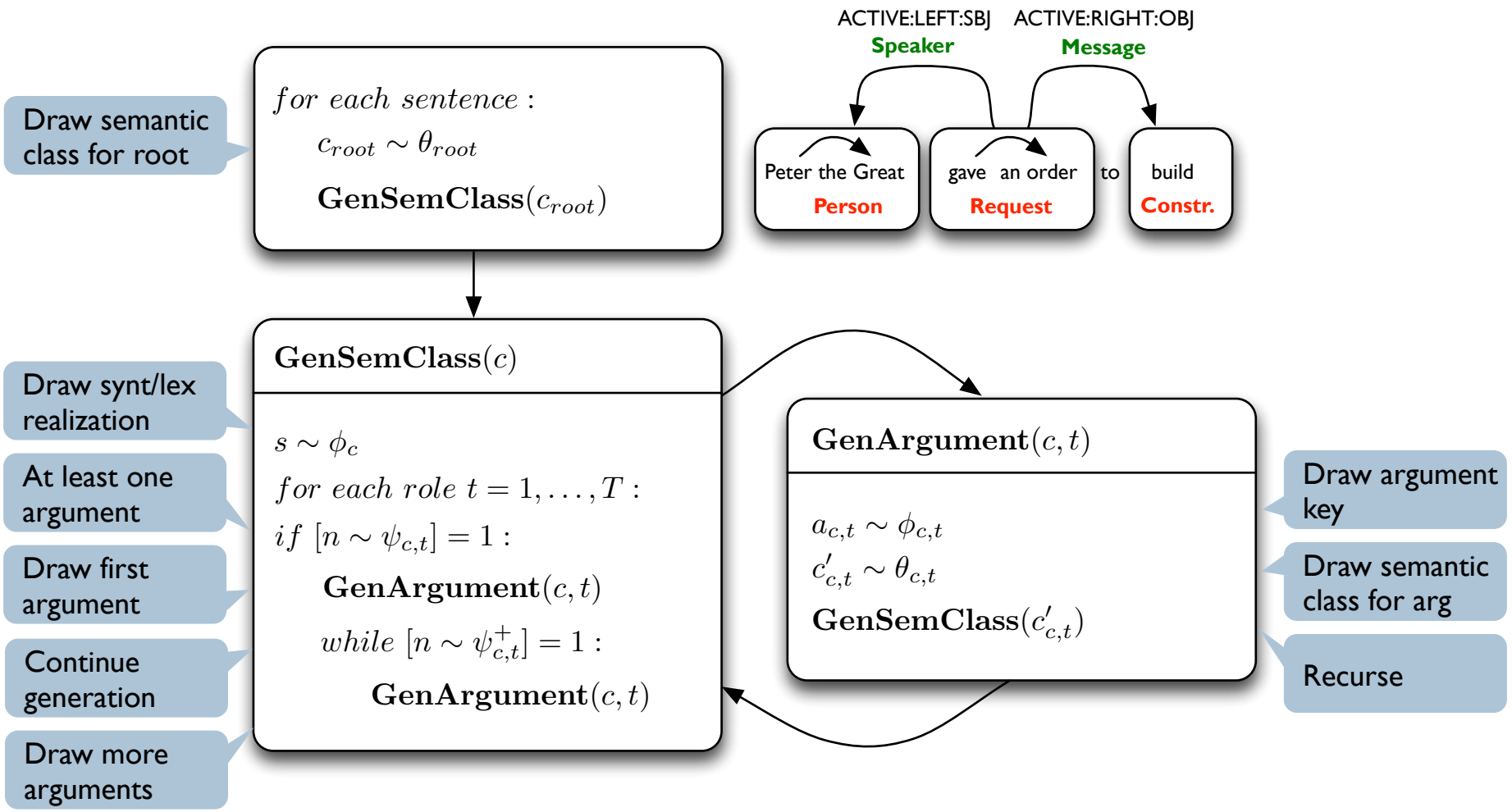
The (Simplified) Model



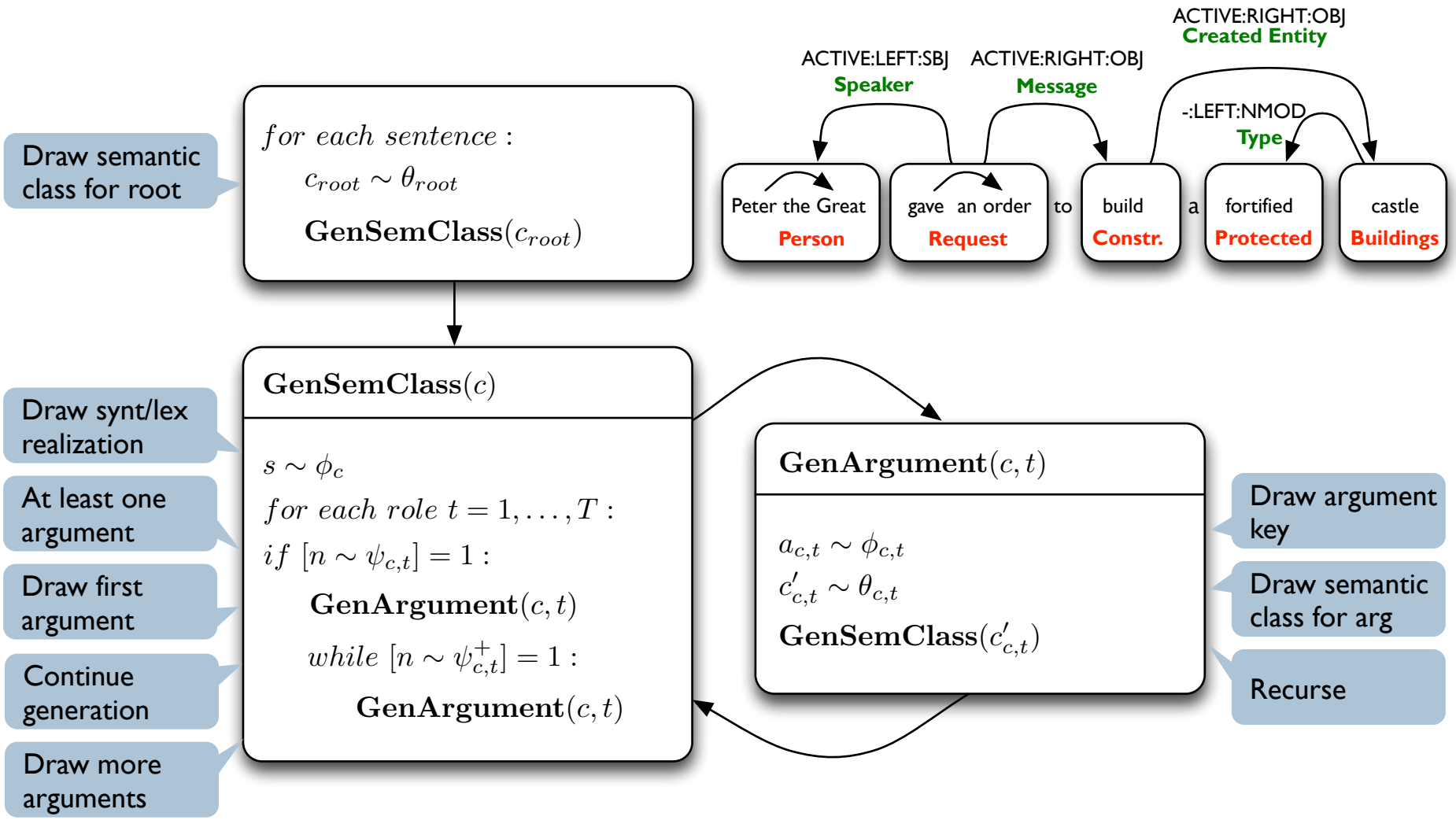
The (Simplified) Model



The (Simplified) Model



The (Simplified) Model



(Titov and Klementiev, ACL 2011, EACL 2012)

Encoding modeling assumptions

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

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

(2) Each semantic class can be verbalized as a sparse distribution over lexemes or syntactic tree fragments

- Sparse priors over syntactic trees (as in Bayesian TSGs [Cohn et al. 07])

(3) Semantically-similar predicates have the same linking between syntax and semantics

- We use sparse Dirichlet priors to encode the linking

(4) The same semantic role rarely appears twice

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

(5) 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

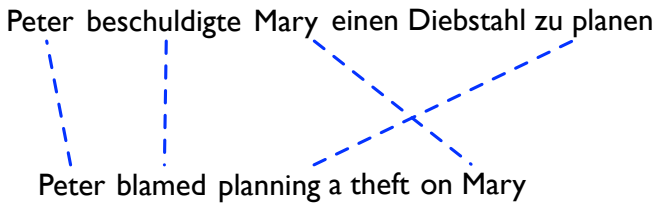
We use MCMC for inference: a form of Metropolis-Hastings Split-Merge sampling

Outline

- Induction of events and their participants
  - unsupervised induction of the frame-semantic representations
  - a Bayesian model and approximate inference
  - cross-lingual extension
  - evaluation (QA, PropBank)
- Induction of semantic representations of words and phrases
  - distributed representations vs. clustering
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, translation scoring)

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

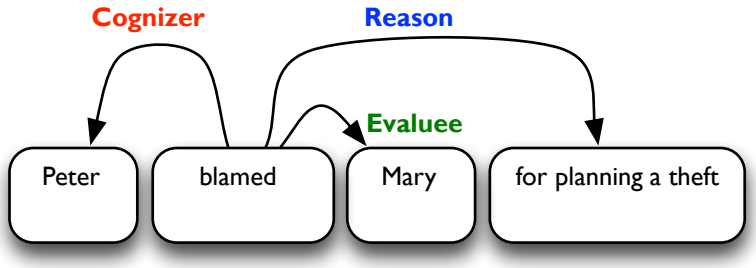


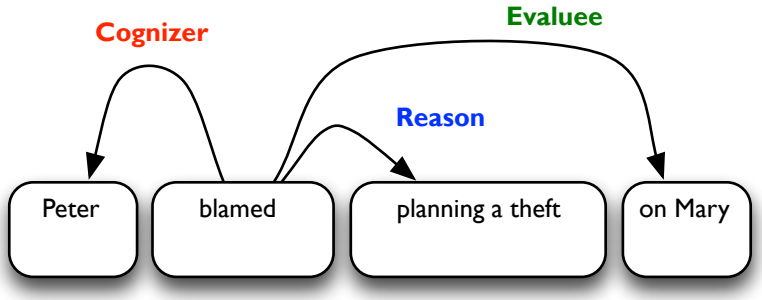
- We use aligned data and induce semantics jointly in multiple languages
  - Only during learning, we apply them to monolingual sentences

(Titov and Klementiev, ACL 2012)

Crosslingual Induction of Semantic Roles

- Consider an example *blame* alternation



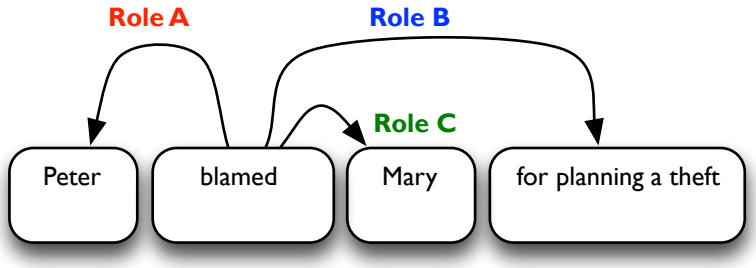


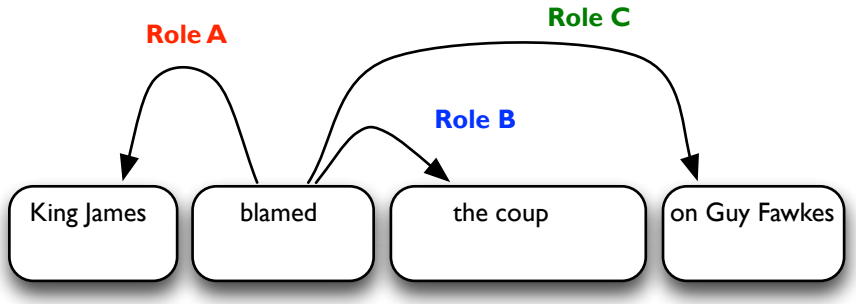
- Learning the corresponding linking is not trivial
  - selectional preferences for all roles are not very restrictive
  - selectional restricutions for Cognizer and Evaluee are overlapping

(Titov and Klementiev, ACL 2012)

Crosslingual Induction of Semantic Roles

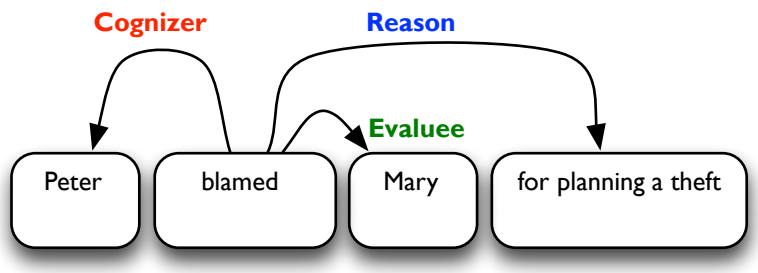
- Consider an example *blame* alternation

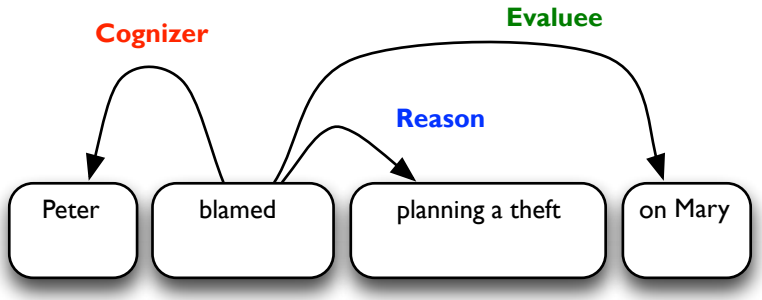




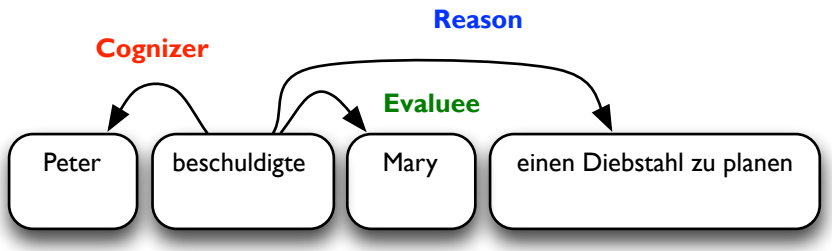
- Learning the corresponding linking is not trivial
  - selectional preferences for all roles are not very restrictive
  - selectional restricutions for Cognizer and Evaluee are overlapping

Crosslingual Induction of Semantic Roles



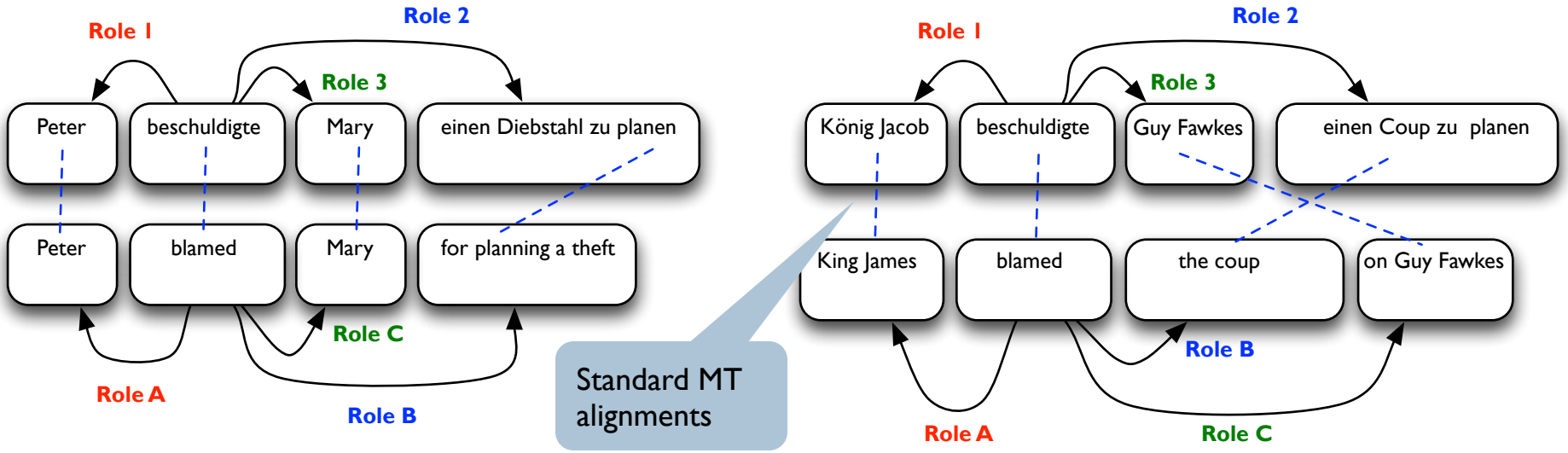


- However, the alternation does not transfer to German
  - Both forms are likely to have the same translation



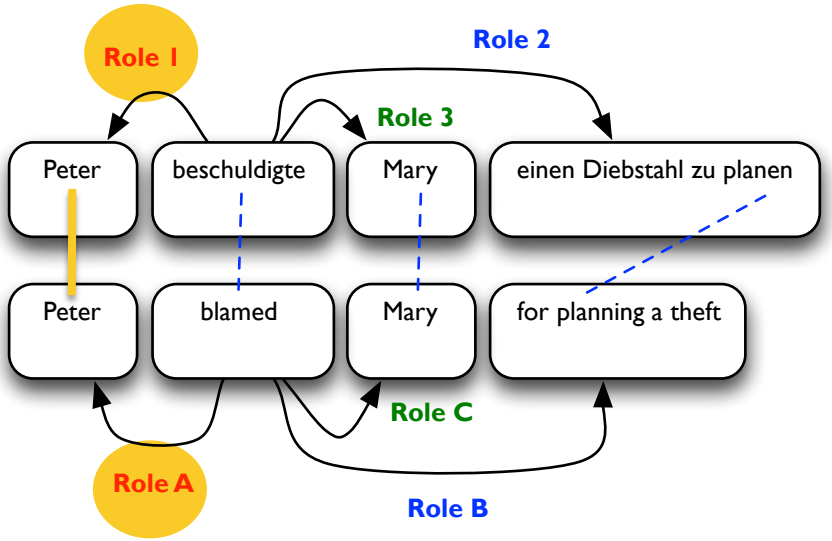
Crosslingual Induction of Semantic Roles

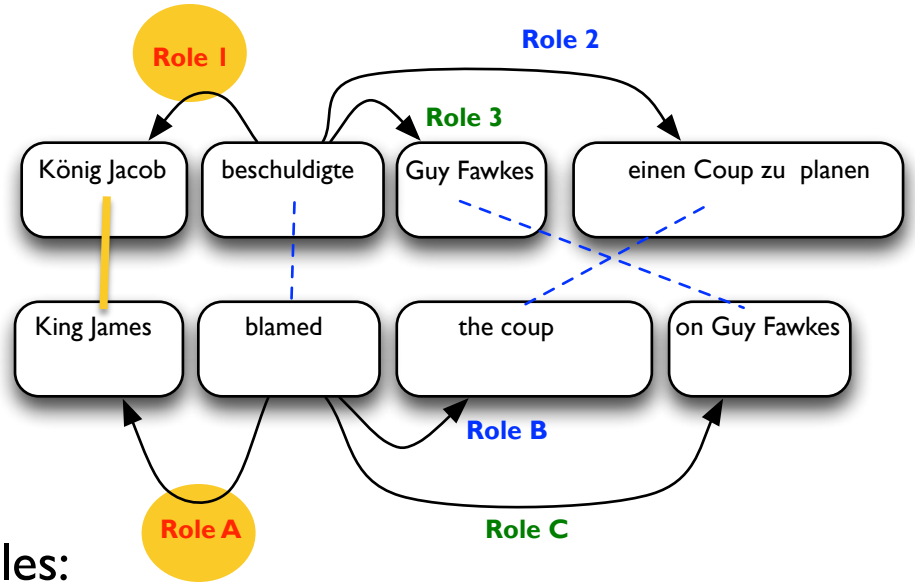
- We want induced roles for aligned sentences to be *consistent*
  - Favoring one-to-one mapping between aligned roles in both languages



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

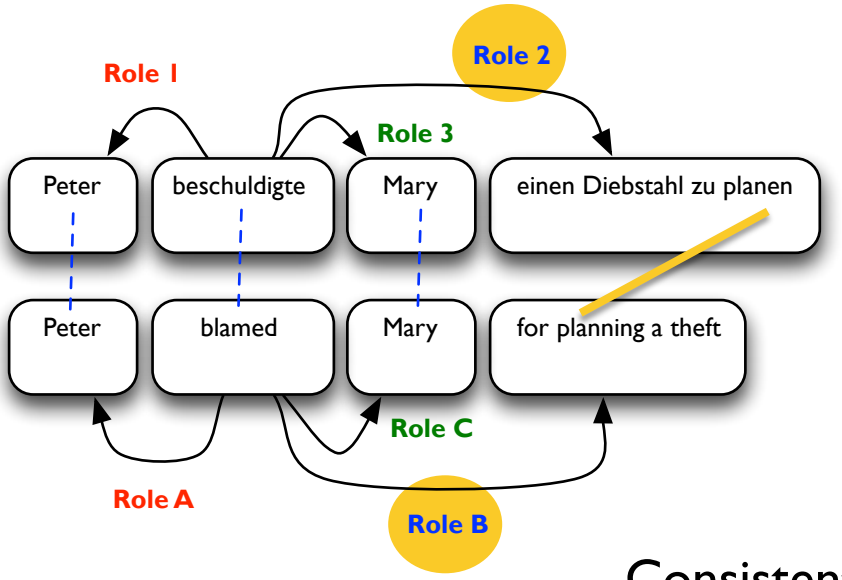


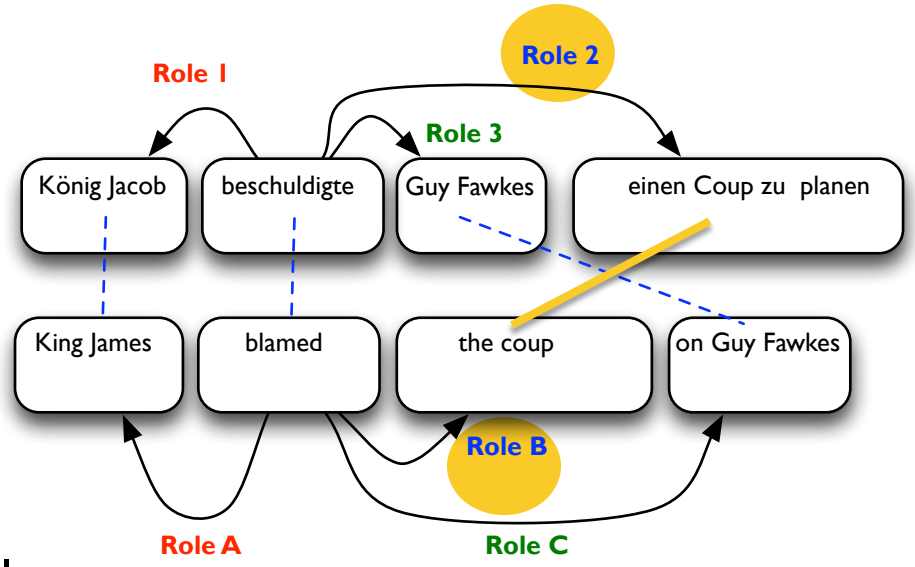


Consistent roles:
A to 1

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

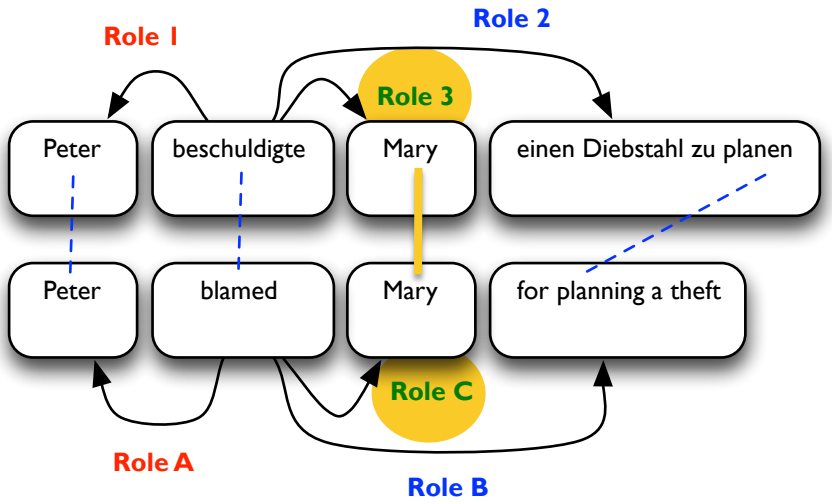


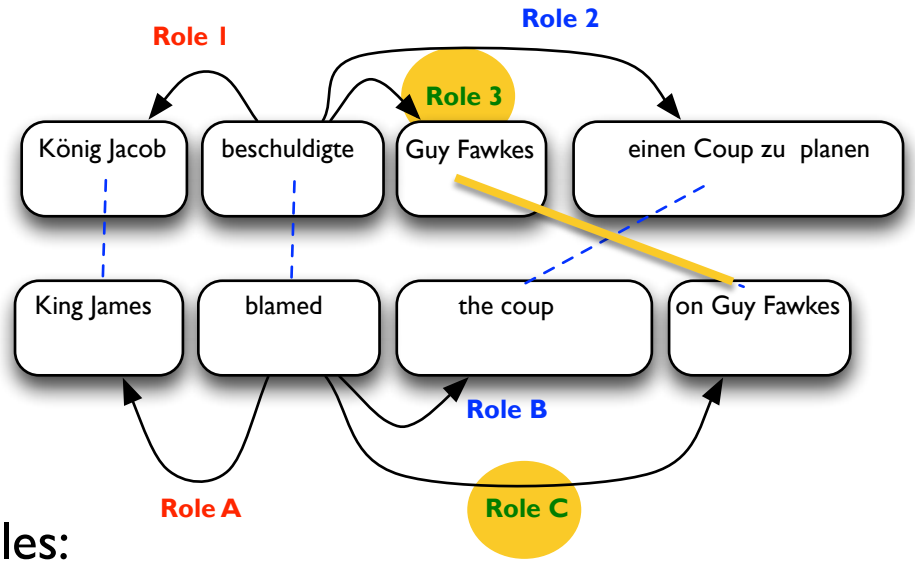


Consistent roles:
A to 1
B to 2

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



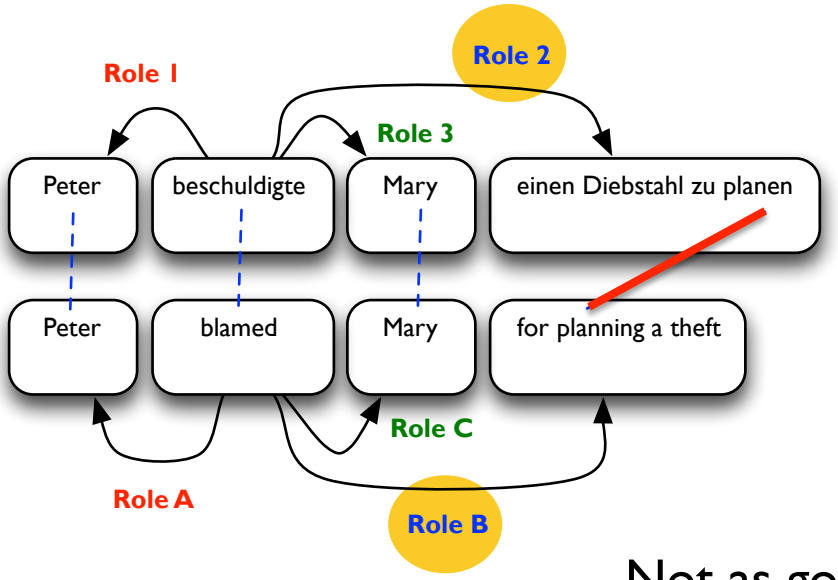


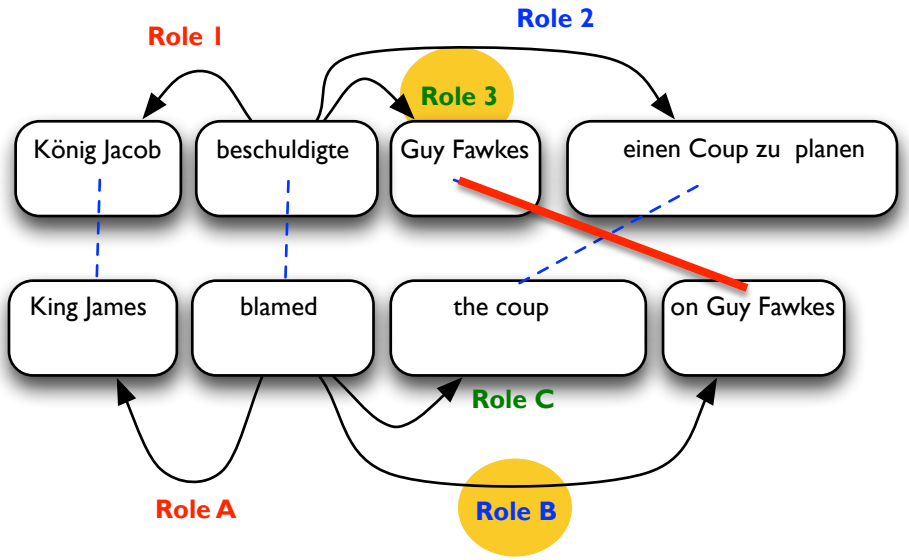
Consistent roles:
A to 1
B to 2
C to 3

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



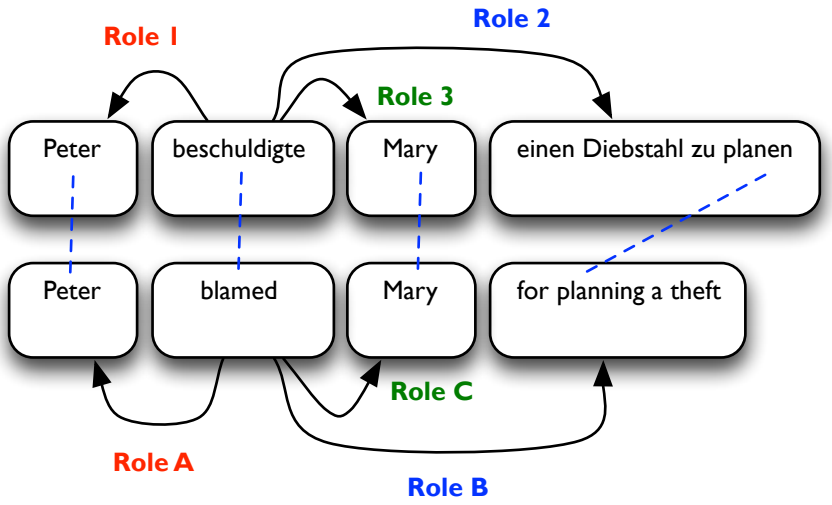


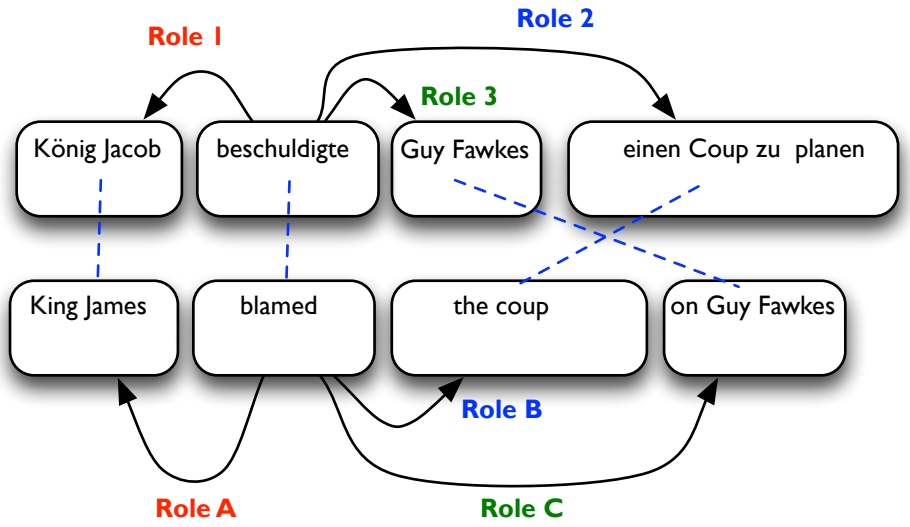
Not as good:
A to 1
B to 2 or 3
C to 3 or 2

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]):
  - a penalty term similar to the expectation criteria [McCallum et al, 08].

Outline

- Induction of events and their participants
  - unsupervised induction of the frame-semantic representations
  - a Bayesian model and approximate inference
  - cross-lingual extension
  - evaluation (QA, PropBank)
- Induction of semantic representations of words and phrases
  - distributed representations vs. clustering
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, translation scoring)

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

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 .

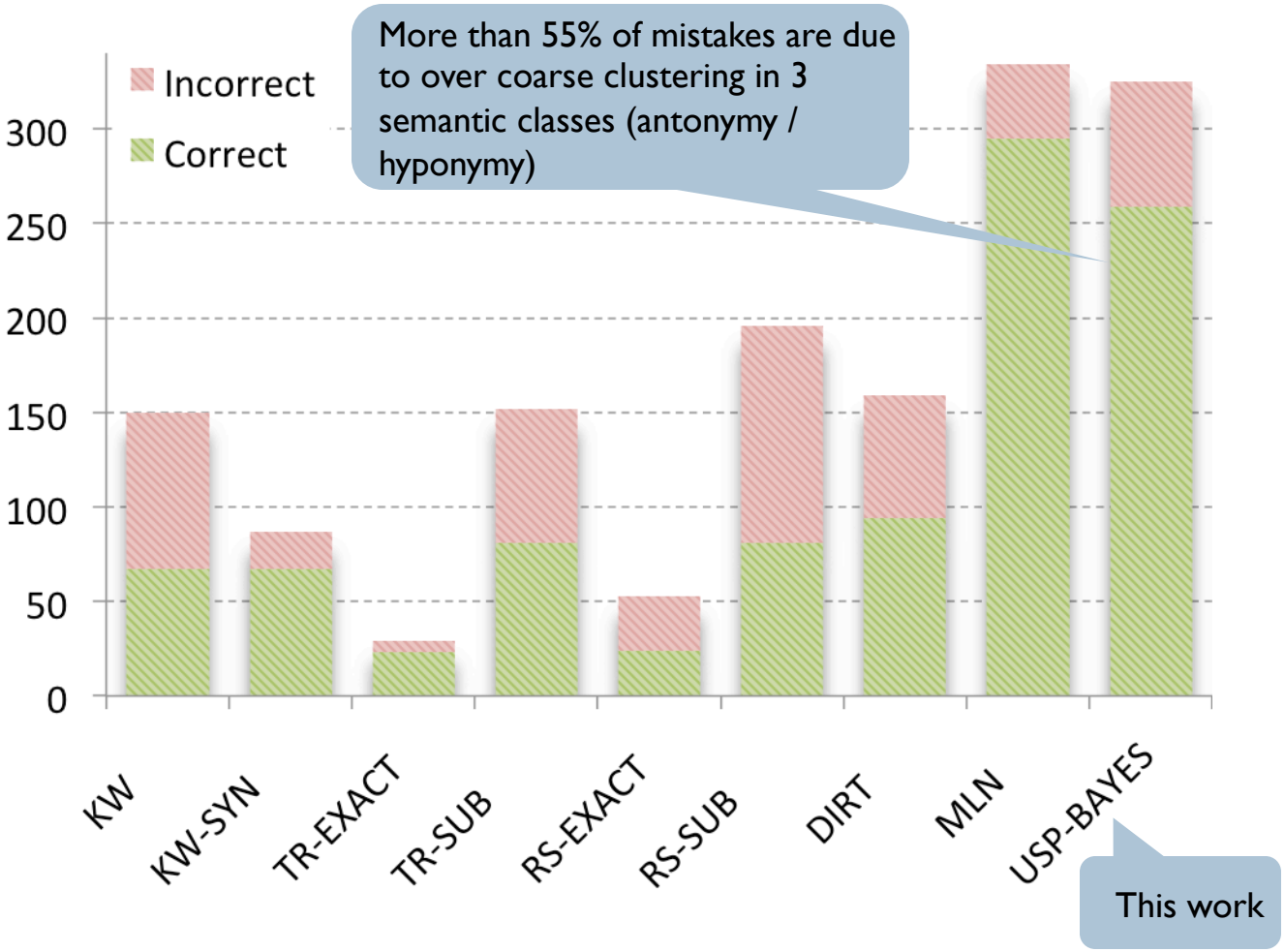
Question: What inhibits tnf-alpha?

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 .

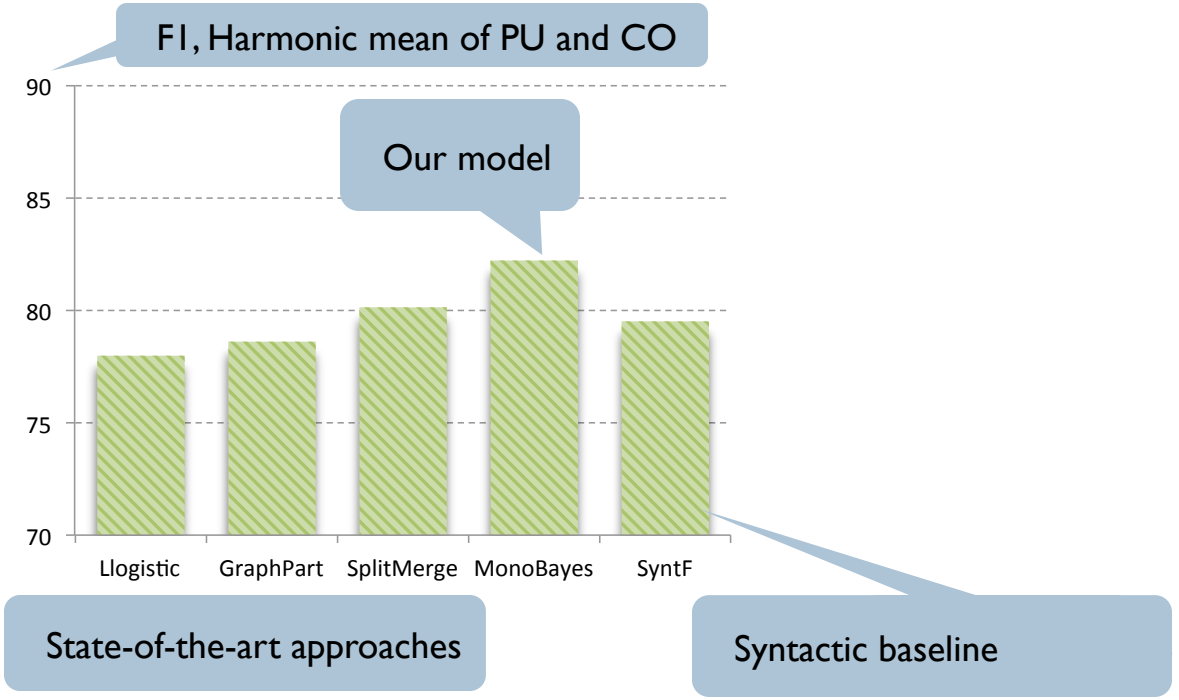
Application-based Evaluation

Question Answering about knowledge in a corpus of biomedical abstracts



Benchmark Dataset: PropBank (CoNLL 08)

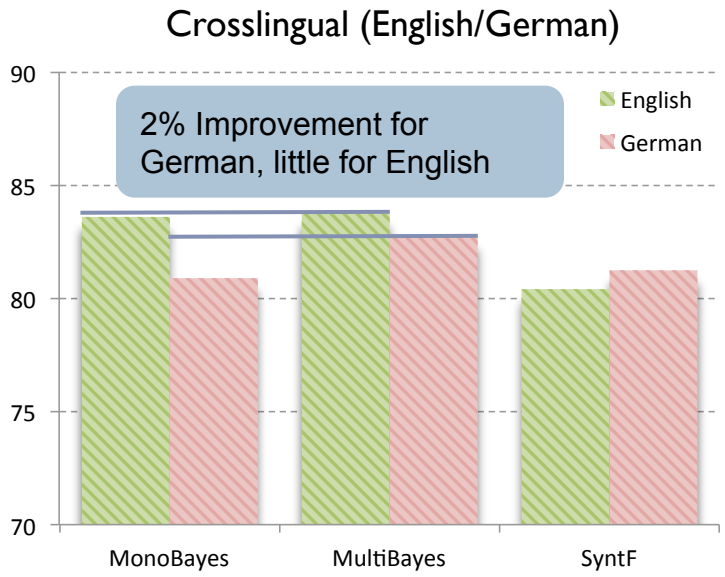
- Evaluation of semantic role induction
- Standard clustering evaluation measures: purity and recall.



See also our results on FrameNet [Modi et al., ILS NAACL 12]

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



Outline

- Induction of events and their participants
  - unsupervised induction of the frame-semantic representations
  - the Bayesian model and approximate inference
  - cross-lingual extension
  - evaluation (QA, PropBank)

- Induction of semantic representations of words and phrases
  - distributed representations vs. clustering
  - cross-lingual induction as multi-task learning
  - evaluation (document classification, phrase-table scoring)

Motivation: Word and Phrase Representations

- NLP systems treating words or phrases as atomic symbols need a lot of annotated data:
  - I.e. vectors with a single one, and many zeros
  - But vocabs are large, many words are rare

Sparsity problems

- Can address this by inducing representations for words instead
  - Use cheap unsupervised data to induce them
  - Use them as features for a learning task

- Very effective on a number of NLP tasks
  - Dependency parsing [Koo et.al., 2008], NER [Turian et.al., 2010],…

Why not clustering as before?

Clustering

- Cluster words into (hierarchical) clusters
- Words defined by cluster prototypes

How to choose granularity?

Many incompatible ways to cluster are often possible

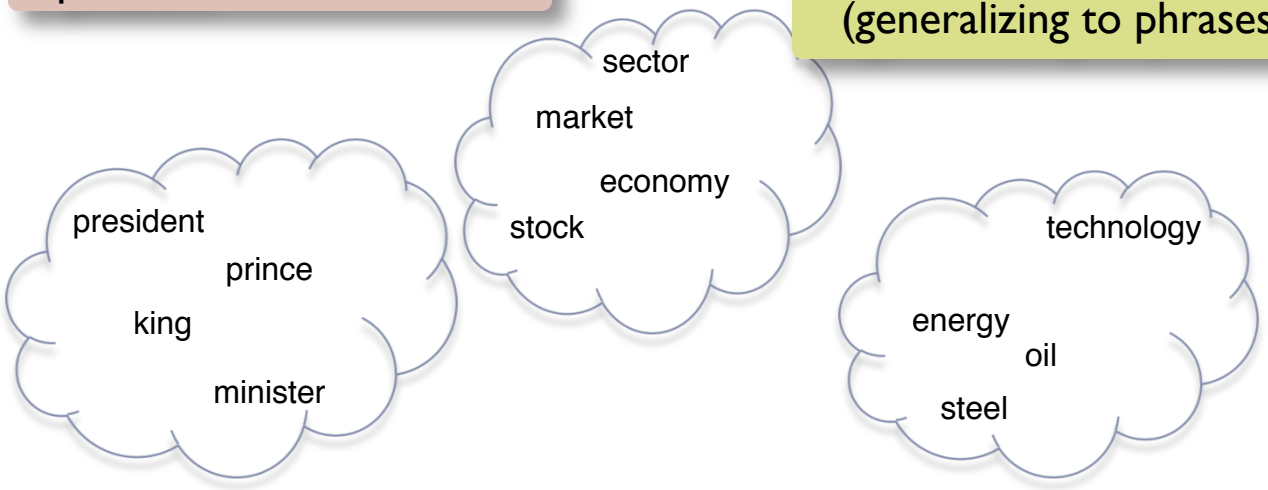
Distributed (= Latent Features)

- Dense embedding

Can encode different levels of granularity

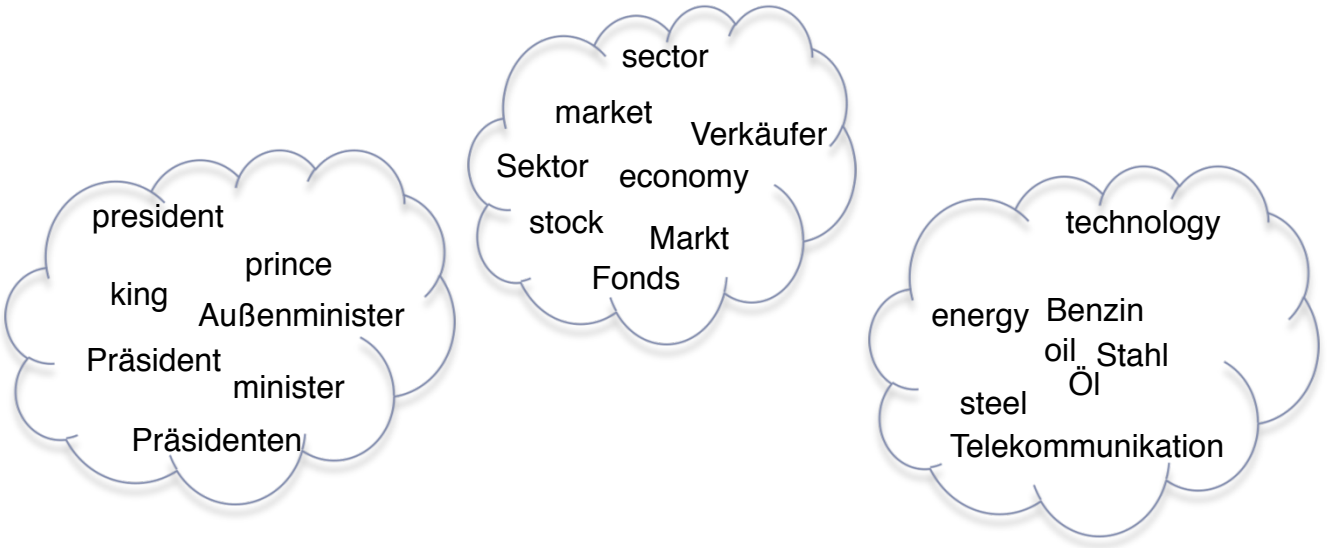
Can encode multiple incompatible clusterings (or multiple senses)

Easier to deal with compositionality (generalizing to phrases)



Why Crosslingual Representations?

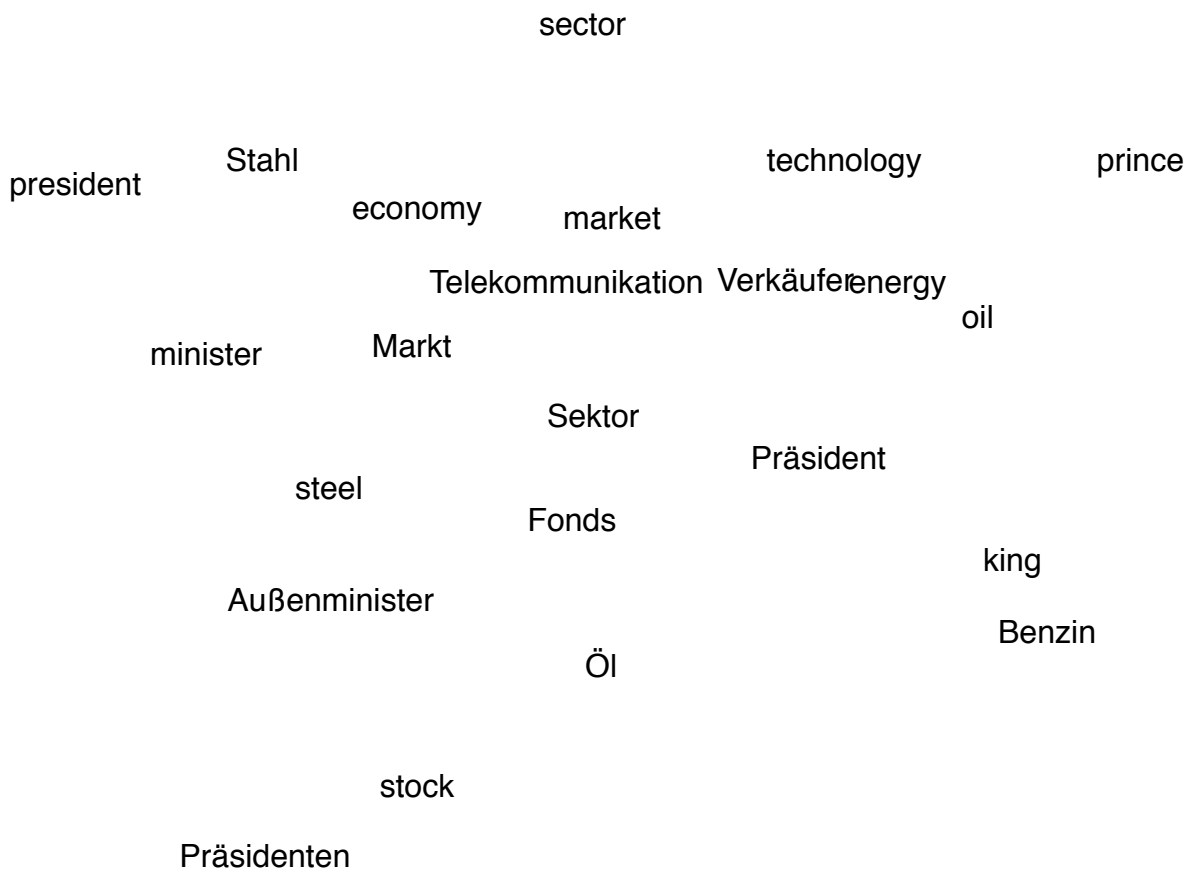
- *Same* representation for both languages:



- Especially important when one of the languages is low resource
  - Learn in one language where annotation is available – apply to the other *directly*!

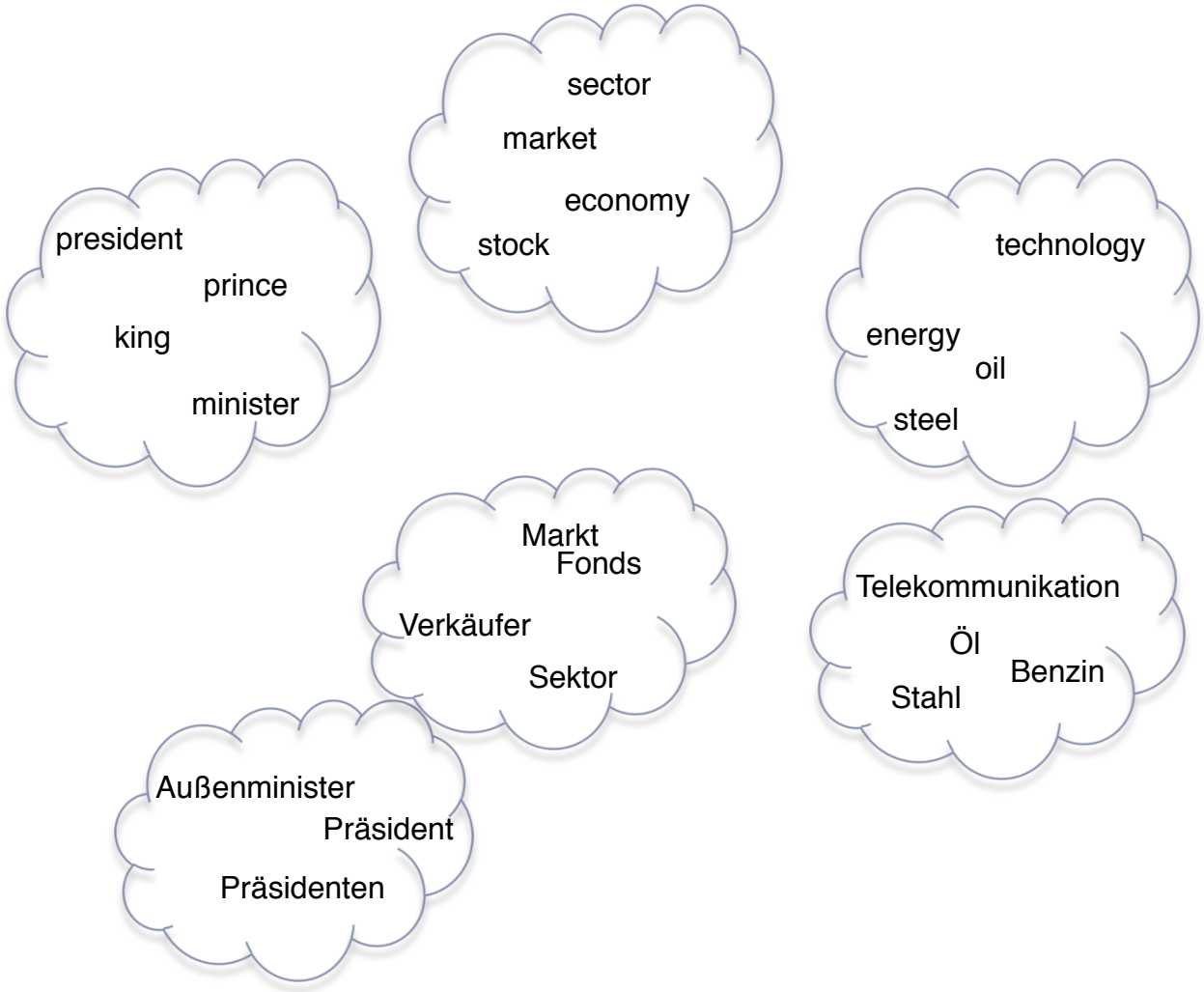
Our work: a general multitask learning inspired framework to induce crosslingual distributed representations

Summary of our Approach



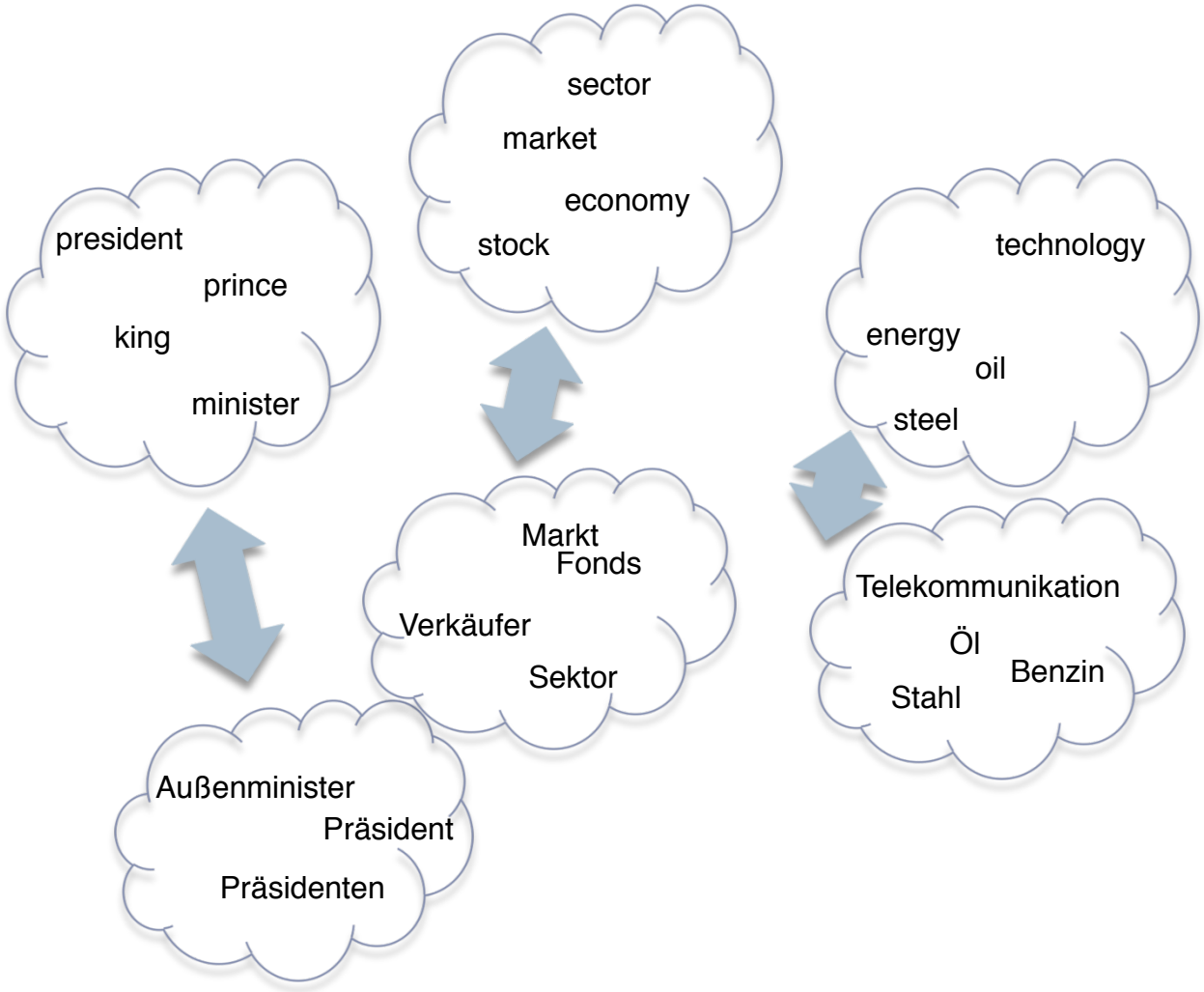
Summary of our Approach

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



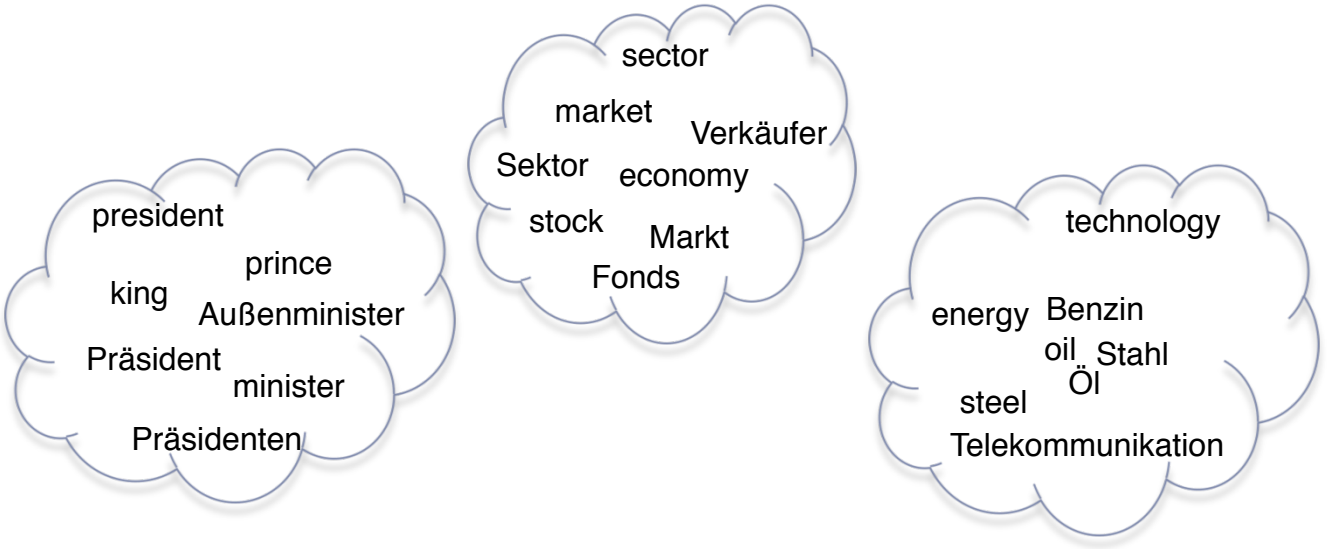
Summary of our Approach

- 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

This work is first to address crosslingual distributed representation induction

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

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

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

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

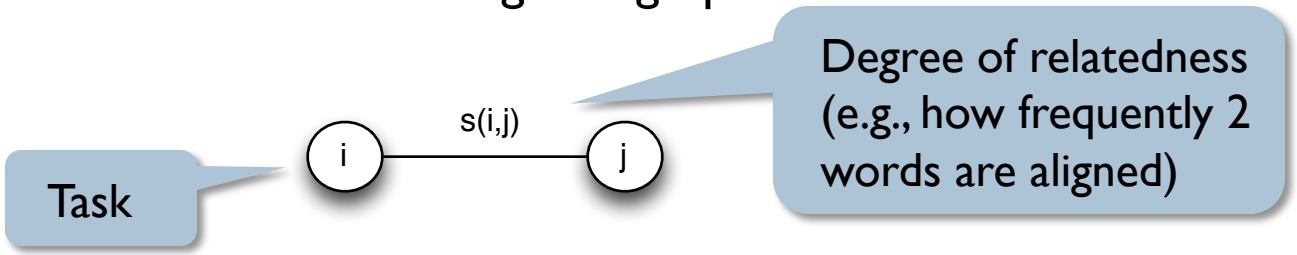
Loss function for a dataset in every language

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)

Evaluation

- Data/Setup
  - Induce 40-dimensional representation of words in German and English
  - RCV1/2 monolingual corpora (~8 million tokens in each language)
  - Europarl parallel data to define the interaction matrix
- Qualitative evaluation
  - Look at a handful of words and their closest neighbors in both languages
- Evaluation on crosslingual document classification
  - Show that the induced representations are informative
  - Evaluated on 4 class classification

Qualitative Evaluation

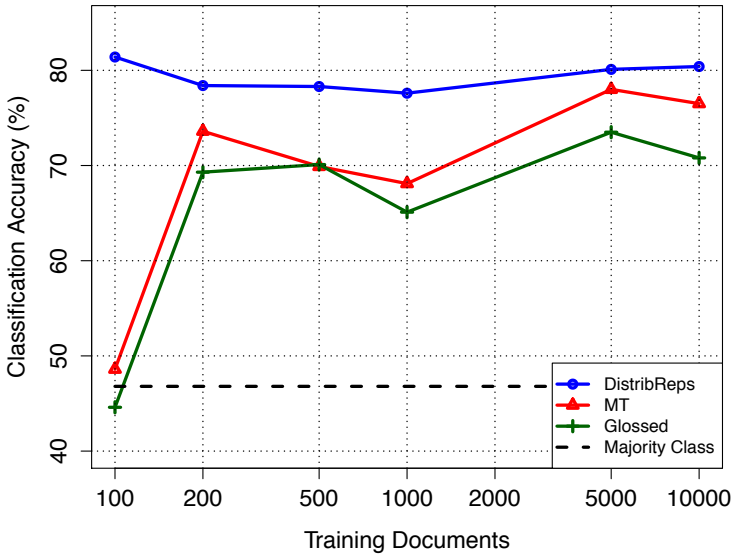
january		*president*		*said*	
en	**de**	**en**	**de**	**en**	**de**
january	januar	president	präsident	said	sagte
february	februar	king	präsidenten	reported	erklärte
november	november	hun	minister	stated	sagten
april	april	areas	staatspräsident	told	meldete
august	august	saddam	hun	declared	berichtete
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

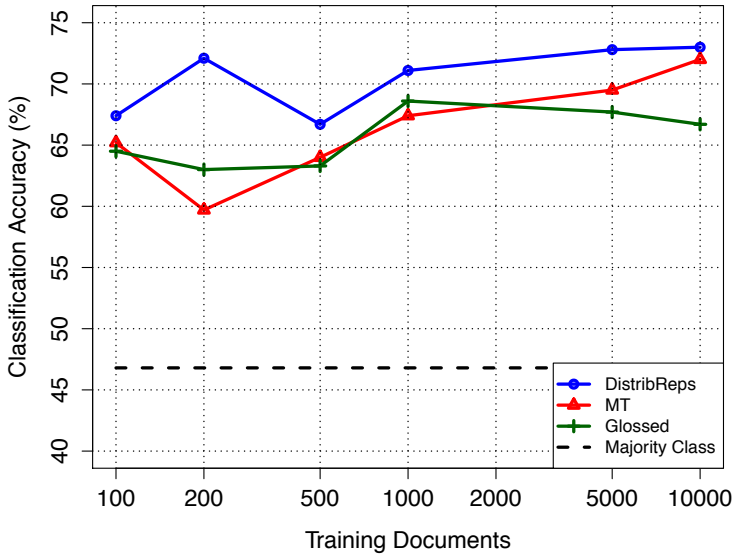
Crosslingual Document Classification

- 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



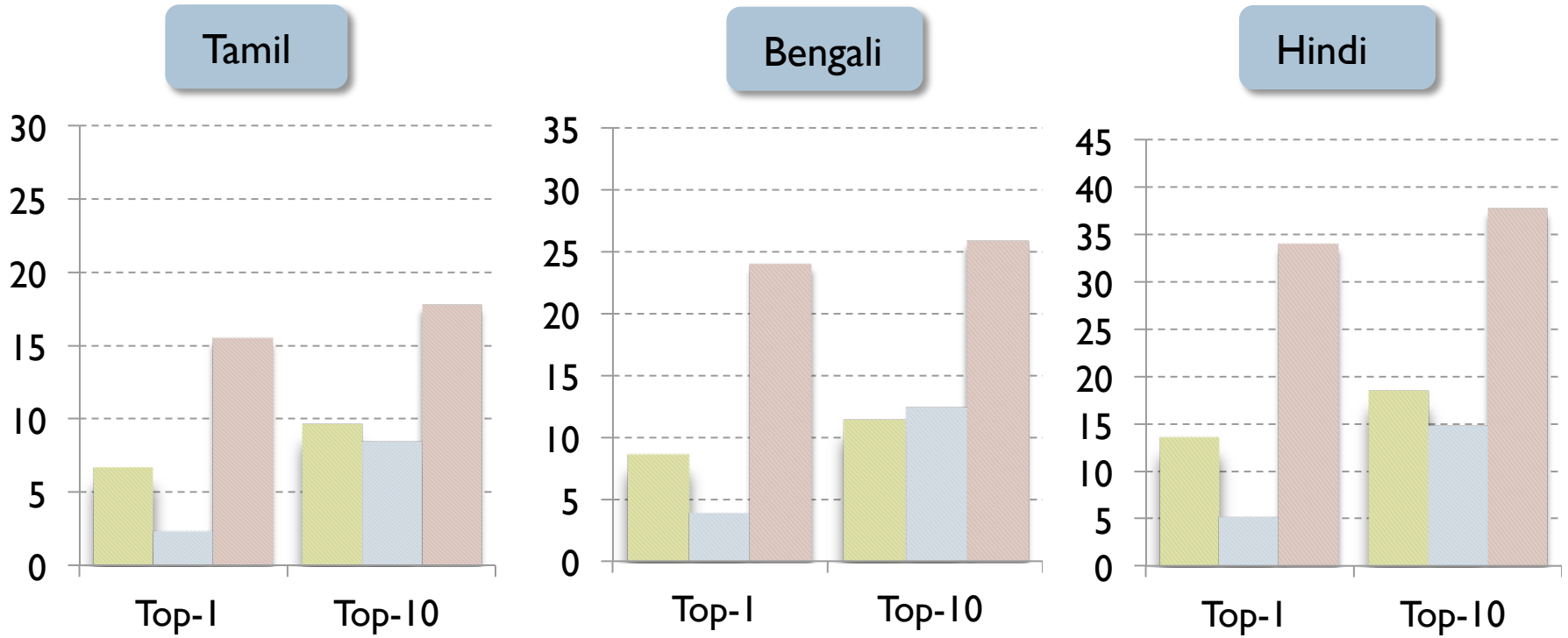
Train: de, Test: en

On-going work: machine translation

- Poor-resource machine translation:
  - 'Little' parallel data (< 1 million words)
  - A lot of monolingual data
- How to proceed?
  - Additional features representing word similarities induced from monolingual data (Klementiev et al., 2012)
- Along with usual MT features, we use features which encode similarity between induced word (and phrase) representations
  - In the simplest form: the Euclidean distance between their representations

On-going work: machine translation

- Results on word translation ranking (aka lexicon induction)
  - Green: scoring from parallel data only
  - Blue: vector space model feature of Klementiev et al (2009)
  - Red: Euclidean distance with crosslingual distributed representations



On-going work: phrases

- In many applications, we would like to have representations of larger linguistic units (e.g., phrases for machine translation)
- How to proceed?
  - We can use models which phrase representations and maybe even phrases themselves (e.g., Socher et al. (2011))
  - The regularization approach does not need to change

Conclusions: cross-lingual representations

	Clustering	Distributed
Words	(Tackstrom et al, 2012)	This talk
Phrases		This talk (?)
Events	This talk	

Similar to relation embedding (e.g., Jenatton et al., NIPS 2012 for the monolingual setting)

- Our data and evaluation scripts are publicly released
  - Code is available by request
- See also our poster at the Deep Learning Workshop tomorrow

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

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

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
- Dative alternation

  I gave the book to Mary

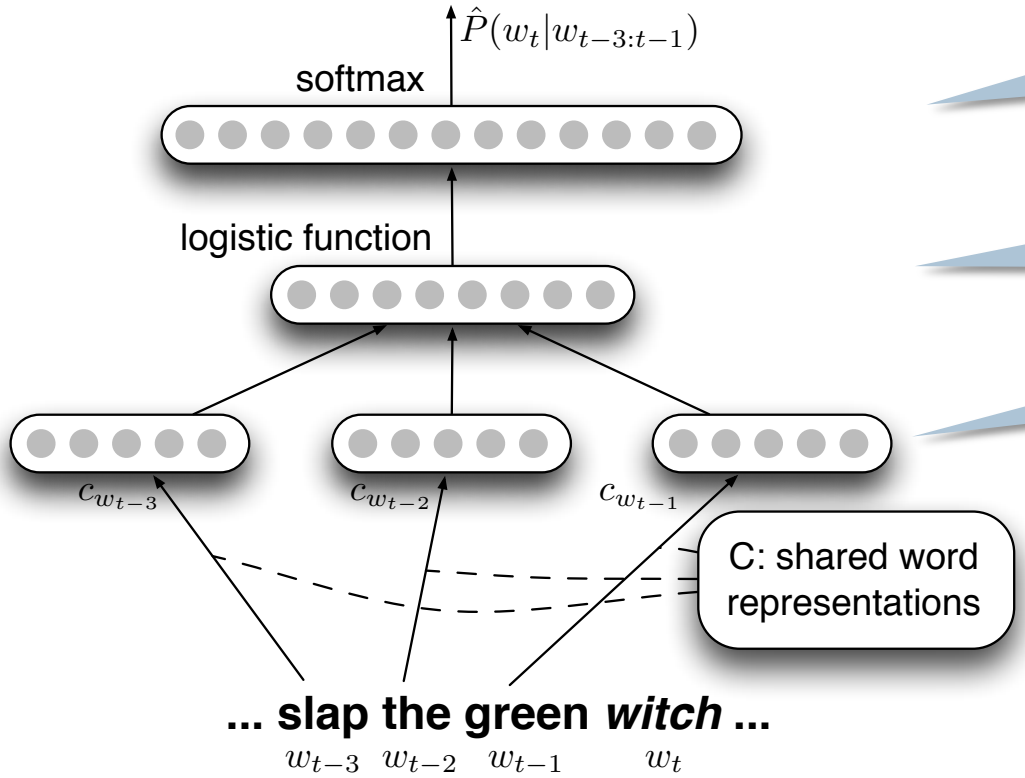
  I gave Mary the book
- Recovery of unnecessary splits introduced by argument keys

Argument Identification

- Assume that sentences are (auto-) annotated with syntactic trees
  - a realistic assumption as syntactic parsers are now available for tens of languages
- Given syntactic representations it is not very hard to identify arguments with reasonable accuracies
  - It can either be handled with simple heuristics based on syntax [Lang and Lapata, 2011]
  - A simple non-lexicalized classifier can be trained on small amount of labeled data
  - Easier for verbs, trickier for nominal predicates

Background: Neural Distributed Representations

Neural probabilistic models learn a latent multi-dimensional representation of words and use them to estimate the probability distribution of word sequences



Turn into prob. distribution (a node for each word)

Apply linear transformation followed by logistic function

Concatenate representations

Map context words to shared representation

Key component!

(Titov and Klementiev, ACL 2011, EACL 2012)

Encoding modeling assumptions

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

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

(2) Each semantic class can be verbalized as a sparse distribution over lexemes or syntactic tree fragments

- Sparse priors over syntactic trees (as in Bayesian TSGs [Cohn et al. 07])

(3) Semantically-similar predicates have the same linking between syntax and semantics

- We use sparse Dirichlet priors to encode the linking

(4) The same semantic role rarely appears twice

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

(5) 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

We use MCMC for inference: a form of Metropolis-Hastings Split-Merge sampling